Reprogramming Protein Synthesis for Cell Engineering

Andrew Vito Anzalone

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Reprogramming Protein Synthesis for Cell Engineering

Andrew Vito Anzalone

Synthetic biology, which aims to enable the design and assembly of customized biological systems, holds great promise for delivering solutions to numerous modern day challenges in agriculture, sustainable energy production, and medicine. However, at its current stage, synthetic biology is not yet equipped with the necessary tools and understanding to reprogram the immensely complex molecular environment of the cell beyond simple proof of concept demonstrations. One current objective within synthetic biology is to create robust tools that can be used to manipulate biological systems in a predictable and reliable manner. While many transcription-based control devices have been reported, little consideration has been given to the eukaryotic protein translation apparatus as a target for engineering gene-regulatory tools.

In this work, we explore the potential for reprogramming the protein synthesis machinery for cell engineering. We begin in **Chapter 1** by reviewing canonical protein synthesis and survey the assortment of translation reprogramming mechanisms that exist in nature, focusing on the role of RNA in these processes. We then cover previous efforts to engineer the protein synthesis machinery and discuss their methodological approaches. Lastly, we examine potential opportunities for engineering protein synthesis that have not yet been explored.

RNA's prominent role in protein synthesis and its amenability to high-throughput *in vitro* selection approaches raises the possibility that the translation apparatus could be

engineered through *in vitro* directed evolution of its RNA components. In **Chapter 2**, we develop an experimental framework for identifying mRNA sequence elements that reprogram protein synthesis, focusing on stop codon readthrough. By adapting a previously developed *in vitro* selection technology called mRNA display, we demonstrate that molecules of RNA derived from expansive libraries of random sequences can be enriched as a result of their translation reprogramming activity. We then analyze these stop codon readthrough signals and propose the use of these sequences for enhanced unnatural amino acid incorporation technologies.

In **Chapter 3**, we apply this very same selection principle for the *in vitro* directed evolution of RNA sequences that stimulate -1 programmed ribosomal frameshifting. Then, using previously reported RNA aptamers, we rationally engineer RNA switches that regulate translation reading frame in response to small molecule inputs. To further optimize switch performance, an *in vivo* directed evolution platform was established. We explore the utility of these RNA switches, particularly their ability to regulate multi-protein stoichiometry, for performing cellular logic operations and controlling cell fate.

A major focus of translation engineering has been the incorporation of unnatural amino acids for fluorescent labeling of proteins in living cells. The successful achievement of this goal will require small molecule fluorophores with desirable biological properties, as well as robust synthetic methods for their production. In **Chapter 4**, we present a scalable approach to oxazine and xanthene fluorophores that utilizes a general diaryl ether synthetic intermediate. Finally, in **Chapter 5**, we describe a photoactivatable oxazine fluorophore and demonstrate its utility as a live-cell imaging reagent with applicability to advanced microscopy techniques.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

5-FOA	5-fluoroorotic acid
A	adenosine
aaRS	aminoacyl tRNA synthetase
bp	base pair
°C	degrees Celsius
C	cytidine
DMF	dimethylformamide
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
E. coli	*Escherichia coli*
et al.	*et alia*
g	gram
G	guanosine
GTP	guanosine triphosphate
gal	galactose
gDNA	genomic DNA
h	hour
L	liter
M	moles per liter
mg	milligram
min	minute

mL	milliliter
nm	nanometer
nM	nanomoles per liter
PCR	polymerase chain reaction
pM	picomoles per liter
raf	raffinose
R	purine, A or G
RNA	ribonucleic acid
rpm	revolutions per minute
S	Svedberg units
s	second
S. cerevisiae	*Saccharomyces cerevisiae*
SC	synthetic complete
T	thymidine
U	uridine
UV	ultraviolet
ug	microgram
uL	microliter
uM	micromoles per liter
Y	pyrimidine, T(U) or C
YPD	yeast peptone dextrose media

Acknowledgments

First, I would like to thank my thesis advisor, Professor Virginia Cornish. I consider myself very fortunate to have had a mentor that is concerned not only with my scientific work, but also my personal well-being and future. She has always kept me and others focused on the bigger picture, and her perspective has been an indispensible resource throughout my graduate career. I think few students receive the opportunity to work in diverse fields during their PhD. I entered the Cornish lab primarily as a chemist, but left with skills and experience in areas spanning molecular biology, ribosome biochemistry, microbial genetics and live cell imaging. I was given the freedom to explore each of these areas, all at my disposal, and had an opportunity to define my own personal research path. This process has been a truly enriching experience, and I thank Professor Cornish for establishing such a nourishing environment for personal and professional growth.

I would also like to thank my graduate committee members, who have been such great resources during my time at Columbia. I thank Professor Ruben Gonzalez, who has always welcomed me to participate in his group's activities (both inside and outside of lab) and has provided me with tremendous feedback over the years. I would also like to thank Dr. Donald Landry and Professor Stephen Goff for their great feedback and support at my committee meetings. Lastly, I would like to thank Professor Peter Sims for graciously offering to serve on my thesis defense committee.

During my time at Columbia, I have had the pleasure of working with an exceptionally talented group of people that created a vibrant and inspiring research environment. I especially thank Rachel Fleisher, dubbed my 'partner in crime', for her support and friendship over the years. I will always have fond memories of our entertaining conversations over coffee. I am extraordinarily grateful to Dr. Nili Ostrov, who taught me almost everything I know about experimental molecular biology and microbiology. Nili has also been, and remains, a phenomenal friend. I am very grateful to Annie Lin, who contributed considerably to the work presented in this thesis and was simply a pleasure to work with. I also thank Zhixing Chen for his energy and enthusiasm,

which has made for many exciting chemistry collaborations. My special thanks go out to all of the other Cornish lab members who I worked with directly on various research projects, including Dr. Josh Avins, Dr. Sonja Billerbeck, Ehud Herbst, Miguel Jiminez, Andy Ng, and Mia Shandell. Lastly, I would like to thank my collaborators outside of the Cornish lab—Liu Wei, for our collaboration on a very exciting Raman imaging project; and Sakellarios Zairis, for our collaboration on sequencing data analysis.

Finally, I am extraordinarily grateful to family for all of their love and support. They have always encouraged me to strive for excellence, and most importantly, believed in me.

To my family, for their love and support.

Chapter 1

Engineering the Translation Machinery

1.0 Chapter outlook

The ribosome and its associated translation components are tasked with coordinating the biosynthesis of the entire cellular proteome. Despite its immense complexity, this ensemble of translation factors collaborates to achieve a strikingly simple program for interpreting the genetic code and synthesizing proteins of specifically defined sequence. The apparent simplicity and logic in the program makes protein translation an extremely attractive and powerful engineering target—merely edit the program and you will alter protein expression. However, the complexity of the translation apparatus should not be overlooked. Hints of underlying intricacies are revealed by the various forms of translation reprogramming found in nature, each of which challenges our simplistic view of the protein synthetic apparatus. While perhaps at first discouraging, great opportunity resides within this complexity, which may be exploited to expand the synthetic and regulatory capabilities of protein synthesis. Here, we briefly review the standard translation program and its key components, and introduce a select set of translation reprogramming mechanisms that have so far been discovered in nature. We then survey the strategies that previous researchers have taken to re-write the translational program, focusing primarily on unnatural amino acid incorporation technologies. Lastly, we examine emerging opportunities for engineering the translation apparatus through previously unexplored approaches.

1.1 The standard translation program

All forms of life (and virus) on earth must use the genetic information stored within their nucleic acid genomes to synthesize proteins, the principal executors of biological function. According to the central dogma of molecular biology[1], the flow of information from nucleic acid to protein is unidirectional, and generally requires that information first be 'transcribed' from a gene's deoxyribonucleic acid (DNA) sequence into ribonucleic acid (RNA). RNA can then be converted into proteins in a process termed 'translation.' Translation is carried out by the ribosome, a highly evolved macromolecular assembly of proteins and RNA composed of a small subunit (30S in bacteria; 40S in eukaryotes) and a large subunit (50S in bacteria; 60S in eukaryotes)[2]. In collaboration with a multitude of translation factors and components, the ribosome coordinates protein biosynthesis according to a well-defined program. This program relies on a set of decoding rules that are applied in succession to non-overlapping triplets of nucleotides, termed codons, within messenger RNA (mRNA)[3].

The code operates with 64 distinct codons, which arise from all possible triplet combinations of the four standard RNA nucleotide monomers: adenosine (A), uridine (U), guanosine (G), and cytidine (C). The program defines each codon to either specify the incorporation of a single amino acid residue (*sense* codons) or the termination of protein synthesis (*stop* or *nonsense* codons). The AUG codon is also used as a signal to initiate protein synthesis. Because proteins are synthesized from 20 standard amino acid building blocks, there is an inherent redundancy in the code such that a single amino acid is often designated by multiple codons. Three of the 64 codons (ochre UAA; opal UGA;

and amber UAG) are typically reserved for programming termination of protein synthesis.

Essentially, the translation program is executed in 3 basic phases: (1) initiation, (2) elongation, and (3) termination (**Fig. 1-1**). Protein synthesis initiates predominantly at AUG codons and requires numerous initiation factors that recruit the requisite translation machinery to prime the apparatus for elongation. Initiation is a highly regulated step, especially in eukaryotes[4], as the decision to initiate generally signifies a commitment to the synthesis of a full-length protein product.

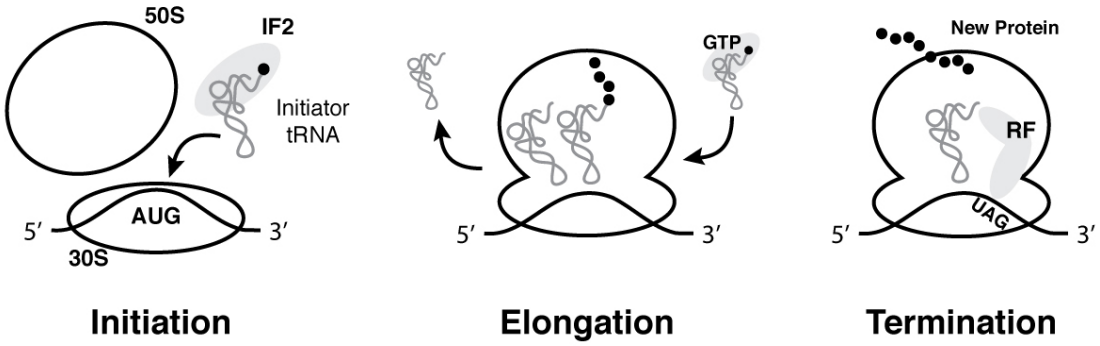


Figure 1-1. Phases of translation. Protein synthesis can be broken down into three basic phases: 1) Initiation requires the cooperation of multiple distinct initiation factors (IFs), the initiator tRNA, and joining of separate 30S and 50S subunits; 2) during elongation, the polypeptide is built through successive decoding and peptide bond forming events; and 3) at termination, the peptide is released from the translation apparatus at stop codons.

During the elongation phase, individual amino acid monomers are delivered to the ribosome in the form of aminoacyl-tRNA, covalently linked through an ester bond to L-shaped transfer-RNA (tRNA) 'adaptor' molecules. In a ternary complex with EF-Tu (eEF1a in eukaryotes) and GTP, aminoacyl-tRNAs rapidly bind to the ribosome and sample the codon in the ribosomal A-site[5]. If the ribosome detects proper (cognate)

pairing between the codon and tRNA anticodon, selection of the aminoacyl-tRNA and accommodation into the A-site are accelerated. Once the aminoacyl-tRNA is properly poised in the A-site, peptide bond formation quickly follows (**Fig. 1-2**)[6].

At this stage of elongation, the entire apparatus must move down the mRNA by three nucleotides, or one codon, to allow for the incorporation of the next amino acid residue. This process, termed translocation, involves capture of a ratcheted, or hybrid, ribosomal state by the GTPase EF-G (EF-2 in eukaryotes)[7,8]. Hydrolysis of GTP by EF-G provides the energetic driving force for translocation of the apparatus, placing the next codon in the ribosome's A-site. The elongation phase continues until the ribosome reaches a stop codon. Here, dedicated class I release factors (RF1 and RF2 in bacteria, and a cooperative complex[9] of eRF1 and eRF3 in eukaryotes) catalyze the hydrolysis of the fully synthesized polypeptide from the P-site tRNA, and recycling factors disassemble the translation apparatus for reuse[10].

The fidelity of protein synthesis is enforced at multiple stages, beginning with aminoacyl-tRNA synthesis by aminoacyl-tRNA synthetase (aaRS) enzymes[11]. These enzymes are responsible for correctly 'charging' free amino acids onto their cognate carrier tRNA molecules[12]. Errors in aminoacylation, which could result in mis-incorporation of an amino acid at an incorrect codon, occur at very low rates ($\sim 1 \times 10^{-5}$) owing to kinetic discrimination and editing by aaRSs[13–15]. Evidence suggests that incorrectly charged tRNAs efficiently incorporate amino acids into the elongating polypeptide[16], suggesting that it is critical that the cellular repertoire of aaRS/tRNA pairs operate orthogonally to one another and with high specificity for their amino acid substrates. The translational apparatus has also evolved to exquisitely discriminate

between codons of similar composition (near-cognate) during the aminoacyl-tRNA selection and accommodation stages of elongation[6]. As a result of kinetic proofreading, error rates for this decoding process, which are codon dependent[17], are estimated to occur between 10^{-3} to 10^{-5} per amino acid incorporation[18].

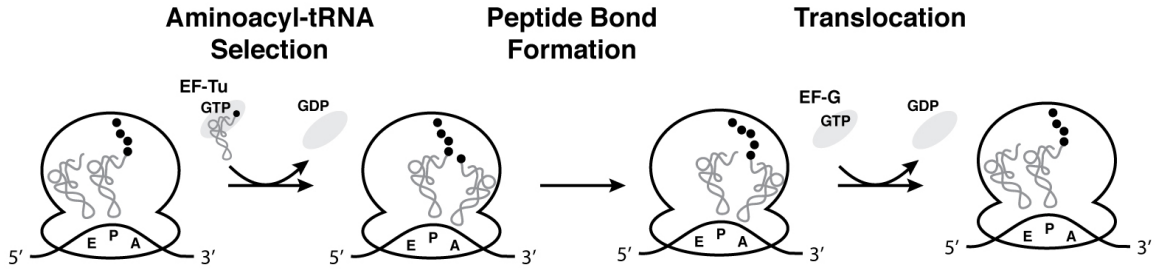


Figure 1-2. The translation elongation cycle. The ribosome contains 3 sites for tRNAs: (i) an A-site, where incoming aminoacyl-tRNAs bind; (ii) a P-site, where the peptidyl tRNA resides; and (iii) and E-site, where deacylated tRNAs exit from the ribosome. There are three basic steps within the elongation cycle. 1) A ternary complex of EF-Tu, aminoacyl-tRNA, and GTP binds to the post-translocation complex. The ribosome selects an aminoacyl-tRNA based on codon-anticodon interactions, then induces EF-Tu to hydrolyze GTP and release tRNA into the ribosomal A-site. 2) Peptide bond formation occurs, transferring the peptide to the A-site tRNA. 3) With the aid of EF-G, the translation apparatus translocates by three nucleotides to allow for a subsequent round of elongation.

Maintenance of translational reading frame is also necessary to ensure that the proper string of codons is translated[19,20]. While mis-incorporation of an amino acid at a single position in a protein could potentially be problematic, it would often only subtly affect protein function. However, a shift in reading frame during elongation changes the sequence of the entire downstream peptide and likely results in the synthesis of a truncated, junk protein product. Errors in translation reading frame maintenance have been estimated at ~ 1 x 10^{-4} in *E. coli*[21]. Maintenance is significantly influenced by the

nature of the codons and tRNAs that are present in the P-site[22], as well as the modification status of the tRNAs[23].

The last phase, termination, must be both efficient and specific for cognate stop codons to prevent unintended extension or truncation of the protein sequence. Unlike sense codons, stop codons are recognized through RNA-protein interactions as opposed to RNA-RNA base-pairing interactions. Similar to sense codon recognition, however, the signal for cognate codon detection must be transmitted to the peptidyl tranferase center to promote specific chemistry. Interestingly, class I release factors resemble tRNAs in shape and size[24]. This structural mimicry is in agreement with their functional similarities. When recognition of a cognate stop codon occurs, the release factor induces an active conformation of the ribosome's peptidyl-transferase center (PTC) and positions a water molecule (using a conserved GGQ motif) for nucleophilic attack on the peptidyl-tRNA ester bond[25]. It has long been recognized that the nucleotide context surrounding the stop codon in the mRNA plays an important role in the efficiency of release at cognate stop codons[26,27]. Therefore, one mechanism to enhance the fidelity of termination could simply be to select for efficient stop codon recognition context[28].

1.2 Translation reprogramming: beyond the standard code

Faithful translation of the code is necessary for ensuring accurate synthesis of the cellular proteome. However, there are surprisingly many instances in which the ribosome deviates from its standard mode of translation[29]. While the classical paradigm of protein synthesis likely predominates for the majority of translation events, examples abound of non-canonical decoding that result in the expression of alternative protein products. Not surprisingly, nature has exploited the inherent complexity and sequence diversity

contained within mRNAs and nascent polypeptide chains to re-wire translation and expand the coding capacity of individual genes. Examples of such re-wiring include programmed stop codon readthrough, programmed ribosomal frameshifting, nascent peptide-mediated stalling, ribosomal 'hopping', and targeted incorporation of non-standard amino acids[30].

1.2.1 Programmed stop codon readthrough in eukaryotes

The amber (UAG), opal (UGA) and ochre (UAA) codons are typically interpreted as termination signals by the eukaryotic translation apparatus, stimulating the hydrolytic release of polypeptides from the ribosome. However, in special cases, these stop codons can be partially reassigned to encode amino acid residues in a context dependent manner. This reprogramming event, termed programmed stop codon readthrough (to be distinguished from stop codon reassignment), has been well documented in plant and animal RNA viruses[31]. Generally, stop codon readthrough serves to enable the synthesis of two distinct proteins from a single mRNA transcript: one, a product of standard translation, and the other, a C-terminally extended product that results from readthrough of the nonsense codon. In RNA viruses, this mechanism is often used to regulate the synthesis of replicase genes[32,33].

Both primary sequences directly adjacent to the stop codon, as well as downstream RNA structures, are capable of promoting this reprogramming event. Primary sequence signals are exemplified by the well characterized Tobacco Mosaic Virus (TMV) readthrough motif[33,34], which has a documented consensus sequence CAR-YYA (R is any purine, and Y is any pyrimidine)—most often CAA-UUA. This hexanucleotide motif stimulates stop codon readthrough with an efficiency of

approximately 5%[35]. The TMV sequence can stimulate readthrough at each of the three stop codons, however it is generally more efficient at the UAG stop codon[27,34]. Other readthrough signals suggest that even a single cytidine nucleotide directly 3'-adjacent to the stop codon is sufficient to induce readthrough[36]. Another class of stop codon readthrough signals comprises a set of downstream structural elements[37], typically taking the form of RNA stem-loops[38] or pseudoknots[39,40] (**Fig. 1-3**). These structures usually begin between 8-9 nucleotides downstream of the last nucleotide of the stop codon. In structure-based readthrough signals in viruses, the sequence of the intervening linker is often conserved and can influence the activity of the readthrough signal[41].

The precise mechanisms by which these reprogramming signals operate are currently unknown and may be diverse in nature. However, it is known that readthrough generally occurs via the selection of near-cognate tRNAs at nonsense codons[29,42]. Therefore, a reasonable hypothesis concerning the overarching principle behind this recoding phenomenon would be that stop codon readthrough signals in some way inhibit release factors from functioning efficiently, thus allowing greater opportunity for selection of near-cognate aminoacyl-tRNAs at nonsense codons[32]. It is also formally possible that these signals operate by accelerating selection of near-cognate aminoacyl-tRNAs through interaction with the ribosome's decoding center.

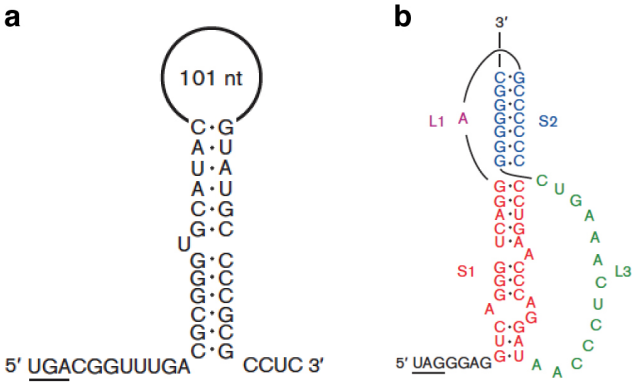


Figure 1-3. RNA secondary structures that induce programmed stop codon readthrough. (**a**) Stem-loop structure of the carnation italian ringspot tombusvirus. (**b**) The murine leukemia virus (MLV) pseudoknot. Figure reproduced from Firth & Brierley[31] and Napthine et al.[38].

Recent cryo-electron microscopy studies[43] of eukaryotic eRF1 bound to the ribosome may provide insight into possible mechanisms of programmed stop codon readthrough. In the published structure, the eRF1 is found interrogating the stop codon, with the two purines at position +2 and +3 stacked (**Fig. 1-4**). Additionally, the 3'-adjacent nucleotide (position +4) is observed being drawn into the ribosomal A-site, stacking with G626 of 18S RNA. This structure is to be contrasted with the structure of bacterial release factor bound to the stop codon, which does not induce this conformational change in the mRNA.

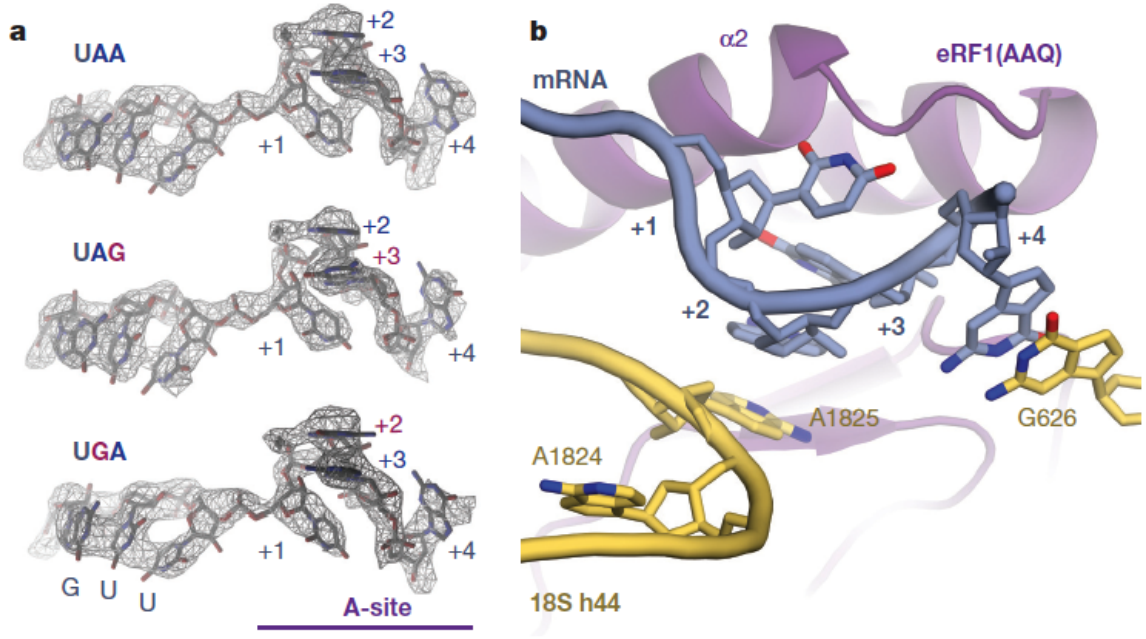


Figure 1-4. Stop codon recognition by eRF1. (**a**) The configuration of the three stop codons (UAA, UAG, and UGA) while being recognized by eRF1. The purine nucleotides in the stop codon (positions +2 and +3) are seen stacking, and the 3'adjacent nucleotide (+4) is pulled into the A-site. (**b**) eRF1 (AAG) inactive mutant bound to the UGA stop codon. The +4 nucleotide is observed stacking with G626 of 18S ribosomal RNA. This figure is reproduced from Brown et al.[43]

One immediate hypothesis regarding the propensity of a +4 position cytidine residue to induce readthrough is apparent from this structure. Base stacking with G626 would presumably favor purine nucleotides at the +4 position, and would be less energetically stabilizing when cytidine occupies the +4 position. Also, the nature of the mRNA distortion may require a 3' region that is malleable to this perturbation. Perhaps, the nature of the downstream nucleotides can influence the ability of the mRNA to tolerate this movement toward the A-site, and in some cases may be unfavorable if it requires the disruption of prev-existing stable interactions. Similarly, structural elements at the entryway of the mRNA tunnel may act to encumber movement of the mRNA toward the ribosomal A-site.

1.2.2 Codon redefinition

The standard genetic code, which utilizes the same 20 amino acids monomers for protein synthesis in all organisms, has long been viewed as universal and unchangeable, the result of a 'frozen accident'[44]. Indeed, the hypothesis put forth by Francis Crick nearly a half-century ago explains many of our observations about the nature of the code and its apparent constancy throughout biology. However, over 20 naturally occurring alternate genetic codes are now known[45], including in mitochondria, refuting this proposition of immutability. The most common changes to the code are in the redefinition of nonsense codons to sense codons. This can be used to specify either a standard amino acid (e.g. UGA coding for tryptophan in mitochondria[46]) or an atypical amino acid such as selenocysteine[47,48] or pyrrolysine[49].

While the former category of codon reclassification (nonsense to standard amino acid) typically applies genome-wide, the redefinition of a stop codon to designate a non-

standard amino acid generally occurs in gene-specific manner. This is especially true for selenocysteine (Sec) incorporation at opal (UGA) stop codons (**Fig. 1-5**), which requires the assistance of a cis-acting RNA motif termed the SECIS (selenocysteine insertion sequence) element for efficient incorporation[50]. While a putative PYLIS element has been proposed[51] to direct pyrrolysine (Pyl) incorporation, the role that this RNA structure plays in promoting pyrrolysine incorporation at amber (UAG) codons is not clear and likely not necessary[52]. Thus, the amber stop codon may be ambiguous in organisms that utilize pyrrolysine.

In contrast, selenocysteine, which is encoded by organisms in all kingdoms of life[53,54], requires several factors for efficienct incorporation at reprogrammed UGA codons. In addition to a downstream SECIS element, Sec incorporation also relies on a dedicated EF-Tu-like elongation factor called SelB (EF-Sec in eukaryotes). In bacteria, this factor is a chimeric protein[55] with one domain that binds specifically to Sec-tRNA^{Sec} with high affinity[56], and a second domain that binds directly to the mRNA SECIS element. In eukaryotes, EF-Sec binds to an adaptor protein called SBP2 that interacts with the SECIS element[57] to recruit Sec-tRNA^{Sec} to the ribosome. Interestingly, Sec is not directly charged onto its tRNA, but instead synthesized from a serine charged intermediate[58,59] Ser-tRNA^{Sec}. While Ser-tRNA^{Sec} would appear to be a viable substrate for translation, the extended acceptor stem of tRNA^{Sec} precludes it from binding to EF-Tu[60,61], and the lack of an anionic Sec residue diminishes the binding affinity to SelB. Thus, Ser-tRNA^{Sec} is prevented from directing serine incorporation at UGA codons.

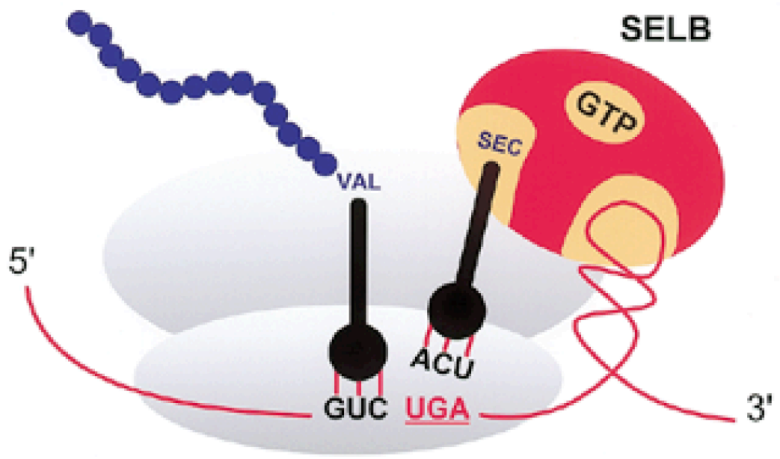


Figure 1-5. SECIS element directed incorporation of selenocysteine at UGA codons. This figure is reproduced from Baranov et al.[62]

1.2.3 -1 programmed ribosomal frameshifting

A common translation reprogramming event found in bacteria and eukaryotic retroviruses is -1 programmed ribosomal frameshifting (-1 PRF)[63]. In eukaryotes, this reprogramming mechanism was first described in Rous sarcoma virus[64,65], where synthesis of the Gag-Pol polyprotein was found to require a ribosomal frameshift event to access the *pol* open reading frame. -1 PRF is now appreciated as a general mechanism employed by RNA viruses to express reverse transcriptase or RNA-dependent RNA-polymerase (replicase). In most retroviruses, including HIV[66], -1 PRF provides a means to strictly control the stoichiometry of Gag and Gag-Pol proteins, which is important for viral fitness[67]. Also, the Gag-Pol fusion enables recruitment of the polymerase into the viral particle by physical connection to a structural component of the virion. In other positive-strand RNA viruses, -1 PRF is also likely important for setting the correct stoichiometry of viral protein synthesis products[68].

As with programmed stop codon readthrough and selenocysteine recoding, -1 PRF requires cis-acting signals within the mRNA to site-specifically trigger the reprogramming event (**Fig. 1-6**). In the case of eukaryotic -1 PRF, two features are necessary: (1) a heptanucleotide slippery site, which is the site of the frameshift event and has the general sequence X-XXY-YYZ (dashes demarcate codons in the original frame; X can be any nucleotide; Y can be A or U; and Z can be A, C or U); and (2) a downstream RNA stimulatory structure, typically a hairpin or pseudoknot[69]. While the slippery site is restricted to sequences that fit the constraints outlined above, the downstream stimulatory elements display a wide range of variability in sequence, size and structure. The diversity found in pseudoknot stimulatory elements[70] suggests that -1 PRF relies less on specific interactions, but rather more general engagement of the translation apparatus.

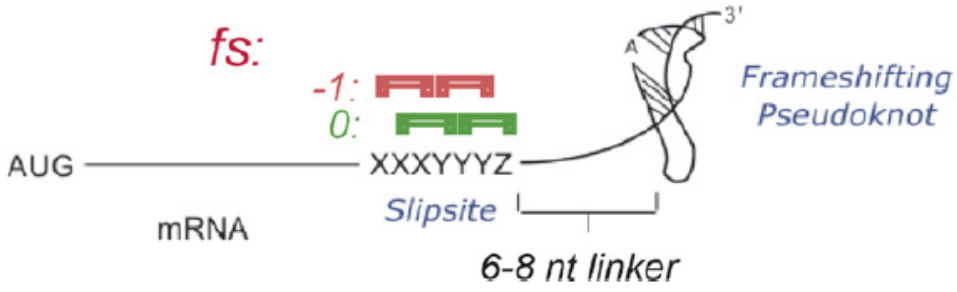


Figure 1-6. Scheme of -1 programmed ribosomal frameshifting. The frameshift even occurs when tRNAs occupy the heptanucleotide slippery site, and is stimulated by a downstream RNA structure (above, an RNA pseudoknot). Figure reproduced from Giedroc & Cornish[70].

Several studies have attempted to unravel the fundamental principles behind ribosomal frameshifting in an effort to construct a descriptive mechanistic model for this important reprogramming event. It was originally hypothesized[71] that the role of the stimulatory structure was to provide mechanical tension for the processing ribosome, and

evidence suggested that the strength of the RNA structure correlated with frameshift activity[72,73]. However, other evidence suggests that the mechanical stability of the RNA structure is not directly correlated to frameshift activity[74], indicating that the operative mechanism may be more subtle.

Recent *in vitro* single molecule fluorescence resonance energy transfer (smFRET) experiments[75,76] have provided some mechanistic insight into this important translation reprogramming event. These studies used the *E. coli* dnaX -1 PRF signal, which is composed of an upstream Shine-Dalgarno sequence in addition to the heptanucleotide slippery site and downstream RNA structure found in eukaryotic -1 PRF signals. In these studies, frameshift signals produced significant ribosome pausing and impeded EF-G catalyzed translocation. Overall, these smFRET experiments suggest that -1 PRF signals stimulate the ribosome to enter a non-standard rotated state that eventually leads to frame slippage, possibly concurrent with aberrant EF-G-mediated translocation.

Another recent study[77] detailed the use of mass spectrometry (MS) and force microscopy to characterize translation products of an mRNA containing the dnaX frameshift signal. Translation products characterized by MS were found to contain peptides that resulted from not just -1 slippage, but also -4 and +2 slippage (accessing the equivalent frame). Additionally, ribosome trajectories corresponding to these positional fluctuations were observed using optical tweezers. The results of this study are intriguing, suggesting that the ribosome makes large movements along the mRNA when encountering a frameshift site, and settles on the -1 frame as a result of acceptable codon-anticodon pairing.

1.3 Reprogramming translation for synthetic biology

The prospect of redefining the intrinsic nature of the genetic code is perhaps one of the most appealing and exciting opportunities in synthetic biology. The ensemble of components that make up the translation apparatus, including the ribosome, tRNAs, mRNAs, aaRS enzymes, and numerous translation factors, is responsible for establishment and maintenance of the standard code. Therefore, efforts to modify the code would logically focus on engineering these translation components. In addition to interpretation of the code, the translation machinery can also control the extent to which different proteins are synthesized, presenting opportunities to regulate gene expression through targeting translation.

1.3.1 Unnatural amino acid mutagenesis

While the standard genetic code utilizes just 20 basic amino acid building blocks, it possesses 64 distinct triplet codons, each of which could theoretically be reserved for a unique purpose. Given the abundance of post-translation modifications, as well as the addition of new amino acids to some genetic codes, a substantial range of useful chemistries are apparently not satisfied by the standard 20 amino acid derivatives. Expansion of the genetic code would therefore allow for the addition of new chemical functionality to the amino acid repertoire, which could be exploited for various applications in basic biology and biotechnology.

The most successful strategy for incorporation of non-standard amino acids has been amber suppressor tRNA technology. Aminoacylation of amber suppressor tRNAs allows for the incorporation of the charged amino acid into proteins in response to the amber codon. This was first demonstrated[78] with unnatural amino acid substrates in an *in*

vitro translation system, where artificially charged tRNAs were prepared by chemical methods (**Fig. 1-7**). This design remains a viable strategy for producing proteins with unnatural side chains, and developments in artificial aminoacylation using evolved ribozymes[79] have assisted in the generation of these unnaturally charged tRNA reagents.

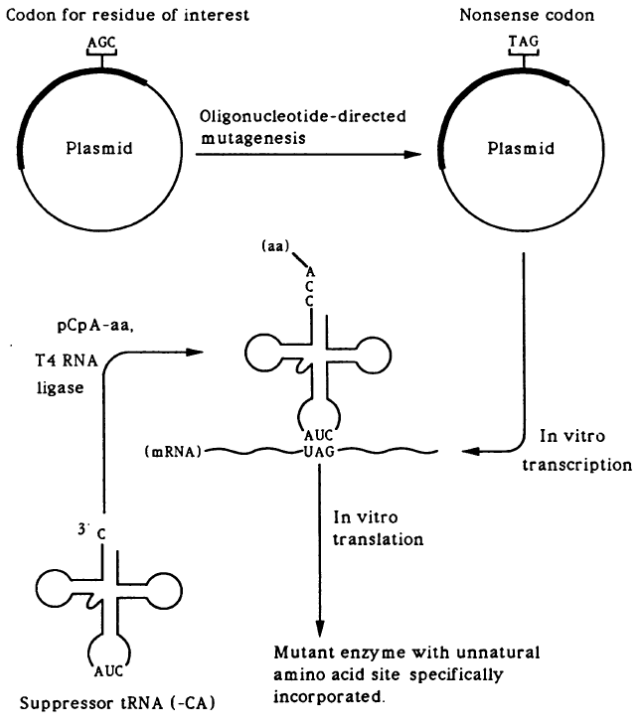


Figure 1-7. Unnatural amino acid incorporation with chemically synthesized suppressor tRNAs. Truncated suppressor tRNAs are *in vitro* transcribed and ligated to aminoacylated dinucleotides. The open reading frame of a protein of interest is mutagenized to encode an amber nonsense codon, then transcribed and translated *in vitro* in the presence of the chemically synthesized tRNA. This figure is reproduced from Noren et al.[78]

A major transformation occurred when Schultz and coworkers reported[80] an evolved aaRS enzyme that could recognize an unnatural amino acid substrate and charge it onto an amber suppressor tRNA in *E. coli*. This required that the imported aaRS/tRNA pair be orthogonal in the host organism to prevent aminoacylation of the suppressor tRNA with natural amino acids. Expression of the evolved orthogonal aaRS/tRNA pair expands the genetic code of *E. coli*, with amber codons partially redefined to specify the non-standard amino acid. Since this initial report, numerous evolved aaRS variants have been reported for genetic encoding of unnatural amino acids in various organisms[81–83].

One valuable application of unnatural amino acid mutagenesis technology is the incorporation of biophysical probes[84] and chemical handles[85] into proteins. Fluorescent amino acids[86–88] have been successfully incorporated into proteins *in vivo*, and recent developments have extended this technology to applications in human cells[89]. A more modular approach has been devised that relies on post-translational modification of the unnatural amino acid using bio-orthogonal click chemistry[90,91]. Recently, inverse Diels-Alder click reactions[92,93] have been exploited for protein specific labeling with fluorescent probes, enabling super-resolution imaging[94].

These and other unnatural amino acid technologies have offered powerful tools for manipulating biological systems. However, widespread adoption of the technology has been limited due to unreliable and low incorporation yields that are highly substrate and protein dependent. Therefore, new approaches are required to enhance the efficiency and specificity of unnatural amino acid mutagenesis if it is to become a truly transformative tool for basic biology and cell engineering.

1.3.2 Engineering translation through RNA

While the protein products of translation are responsible for carrying out most cellular functions, the process of protein synthesis itself, which is so central to life, conspicuously employs another biopolymer—RNA. From the coding messenger RNA template, to the tRNA 'adaptor' molecules, to the decoding and peptidyl-transferase centers of the ribosome, RNA's role in protein translation is paramount. RNA is also widely used by nature to 'edit' the standard translation program and alter the outcome of protein synthesis. Thus, any attempt to modify or engineer protein translation would wisely consider its RNA components as targets.

Many efforts to engineer the RNA components of translation began with tRNAs[95,96]. tRNAs are attractive options for engineering, as they represent relatively short stretches of RNA sequence that are well understood in terms of structure and function. Engineered quadruplet decoding tRNAs, which suppress frameshift mutations as opposed to nonsense codons, have been used for unnatural amino acid incorporation[97]. Orthogonal tRNA/aaRS pairs have been generated by mutation of the tRNA body followed by complementary changes in the partner aaRS[98,99]. Additionally, the anticodon stem of a an amber suppressing phosphoseryl-tRNA has been optimized for high suppression efficiency by *in vivo* directed evolution[100]. While recent efforts to engineer tRNAs have taken more ambition approaches, dramatic changes to the tRNA body have not been extensively explored. For example, $tRNA^{Sec}$, which possesses an extended acceptor arm unlike any other natural tRNA[101], demonstrates the possibilities for engineering non-canonical tRNA variants with specialized function.

The mRNA also has great potential as a component for translation engineering, especially considering the numerous translation reprogramming and recoding events that utilize signals contained within the mRNA. However, with the exception of internal ribosome entry sites (IRES)[102], few of these mechanisms have been adopted for synthetic biology. A great deal of research has been devoted to developing unnatural amino acid mutagenesis technologies using stop codon suppression. Yet, none of these approaches have taken advantage of the natural design of selenocysteine incorporation. Moreover, other reprogramming mechanisms such as stop codon readthrough and -1 PRF have potential applications for gene regulation through engineering of their RNA components.

Adaptation of these and other mRNA-based translation control mechanisms should prove fruitful for future cell engineering efforts.

Without question, the most ambitious RNA target for translation engineering is the ribosome. High-resolution structures[103] have recently made the rational alteration of ribosomal RNA a realistic endeavor. One intriguing example has been the development of orthogonal bacterial ribosome-mRNA pairs[104], which were created by evolution of the Shine-Dalgarno and anti-Shine-Dalgarno sequences in the mRNA and ribosome, respectively. These orthogonal ribosomes were further evolved to more efficiently decode quadruplet codons by mutation of 16S RNA in the small (30S) ribosomal subunit[105]. Recently, bacterial ribosomes with tethered 30S and 50S subunits were reported[106,107], enabling mutagenesis of the large ribosomal subunit with genetically encoded orthogonal ribosomes. These developments open the opportunity to refashion the peptidyl-transferase center to promote new forms of catalysis, introducing perhaps the grandest engineering challenge of all—changing the very nature of the chemistry of translation.

1.4 Conclusions

The protein synthesis machinery is a complex assembly of RNA and protein components that collaborate to translate the genetic code. Nature provides us with several examples of ways in which the ribosome can be reprogrammed in a gene-specific manner to achieve expanded synthetic and regulatory capabilities. Modification of the standard genetic code holds enormous potential for engineering biological systems, and we would argue for a biomimetic approach that exploits nature's solutions to these engineering challenges.

1.5 References

1. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).
2. Steitz, T. A. A structural understanding of the dynamic ribosome machine. *Nat. Rev. Mol. Cell Biol.* **9,** 242–253 (2008).
3. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).
4. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11,** 113–127 (2010).
5. Pape, T., Wintermeyer, W. & Rodnina, M. V. Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E. coli ribosome. *EMBO J.* **17,** 7490–7497 (1998).
6. Rodnina, M. V. & Wintermeyer, W. Fidelity of Aminoacyl-tRNA Selection on the Ribosome: Kinetic and Structural Mechanisms. *Annu. Rev. Biochem.* **70,** 415–435 (2001).
7. Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez Jr., R. L. Coupling of Ribosomal L1 Stalk and tRNA Dynamics during Translation Elongation. *Mol. Cell* **30,** 348–359 (2008).
8. Tourigny, D. S., Fernández, I. S., Kelley, A. C. & Ramakrishnan, V. Elongation Factor G Bound to the Ribosome in an Intermediate State of Translocation. *Science* **340,** 1235490 (2013).
9. Stansfield, I. *et al.* The products of the SUP45 (eRF1) and SUP35 genes interact to mediate translation termination in Saccharomyces cerevisiae. *EMBO J.* **14,** 4365–4373 (1995).
10. Dever, T. E. & Green, R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4,** a013706 (2012).
11. Carter, C. W. Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. *Annu. Rev. Biochem.* **62,** 715–748 (1993).
12. Ibba, M. & Söll, D. Aminoacyl-tRNA Synthesis. *Annu. Rev. Biochem.* **69,** 617–650 (2000).

13. Baldwin, A. N. & Berg, P. Transfer ribonucleic acid-induced hydrolysis of valyladenylate bound to isoleucyl ribonucleic acid synthetase. *J. Biol. Chem.* **241,** 839–845 (1966).
14. Eldred, E. W. & Schimmel, P. R. Rapid deacylation by isoleucyl transfer ribonucleic acid synthetase of isoleucine-specific transfer ribonucleic acid aminoacylated with valine. *J. Biol. Chem.* **247,** 2961–2964 (1972).
15. Jakubowski, H. & Goldman, E. Editing of errors in selection of amino acids for protein synthesis. *Microbiol. Rev.* **56,** 412–429 (1992).
16. Effraim, P. R. *et al.* Natural amino acids do not require their native tRNAs for efficient selection by the ribosome. *Nat. Chem. Biol.* **5,** 947–953 (2009).
17. Kramer, E. B., Vallabhaneni, H., Mayer, L. M. & Farabaugh, P. J. A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA* **16,** 1797–1808 (2010).
18. Zaher, H. S. & Green, R. Fidelity at the Molecular Level: Lessons from Protein Synthesis. *Cell* **136,** 746–762 (2009).
19. Farabaugh, P. J. Translational frameshifting: implications for the mechanism of translational frame maintenance. *Prog. Nucleic Acid Res. Mol. Biol.* **64,** 131–170 (2000).
20. Drummond, D. A. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10,** 715–724 (2009).
21. Jørgensen, F. & Kurland, C. G. Processivity errors of gene expression in Escherichia coli. *J. Mol. Biol.* **215,** 511–521 (1990).
22. Jenner, L. B., Demeshkina, N., Yusupova, G. & Yusupov, M. Structural aspects of messenger RNA reading frame maintenance by the ribosome. *Nat. Struct. Mol. Biol.* **17,** 555–560 (2010).
23. Urbonavicius, J., Qian, Q., Durand, J. M., Hagervall, T. G. & Björk, G. R. Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J.* **20,** 4863–4873 (2001).
24. Moffat, J. G. & Tate, W. P. A single proteolytic cleavage in release factor 2 stabilizes ribosome binding and abolishes peptidyl-tRNA hydrolysis activity. *J. Biol. Chem.* **269,** 18899–18903 (1994).

25. Youngman, E. M., McDonald, M. E. & Green, R. Peptide Release on the Ribosome: Mechanism and Implications for Translational Control. *Annu. Rev. Microbiol.* **62,** 353–373 (2008).
26. Miller, J. H. & Albertini, A. M. Effects of surrounding sequence on the suppression of nonsense codons. *J. Mol. Biol.* **164,** 59–71 (1983).
27. Bonetti, B., Fu, L., Moon, J. & Bedwell, D. M. The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. *J. Mol. Biol.* **251,** 334–345 (1995).
28. Brown, C. M., Stockwell, P. A., Trotman, C. N. & Tate, W. P. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Res.* **18,** 6339–6345 (1990).
29. *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. (Springer New York, 2010). at <http://link.springer.com/10.1007/978-0-387-89382-2>
30. Baranov, P. V., Atkins, J. F. & Yordanova, M. M. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nat. Rev. Genet.* **16,** 517–529 (2015).
31. Firth, A. E. & Brierley, I. Non-canonical translation in RNA viruses. *J. Gen. Virol.* **93,** 1385–1409 (2012).
32. Bertram, G., Innes, S., Minella, O., Richardson, J. & Stansfield, I. Endless possibilities: translation termination and stop codon recognition. *Microbiol. Read. Engl.* **147,** 255–269 (2001).
33. Pelham, H. R. Leaky UAG termination codon in tobacco mosaic virus RNA. *Nature* **272,** 469–471 (1978).
34. Skuzeski, J. M., Nichols, L. M., Gesteland, R. F. & Atkins, J. F. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J. Mol. Biol.* **218,** 365–373 (1991).
35. Dreher, T. W. & Miller, W. A. Translational control in positive strand RNA plant viruses. *Virology* **344,** 185–197 (2006).

36. Li, G. & Rice, C. M. The signal for translational readthrough of a UGA codon in Sindbis virus RNA involves a single cytidine residue immediately downstream of the termination codon. *J. Virol.* **67,** 5062–5067 (1993).
37. Firth, A. E., Wills, N. M., Gesteland, R. F. & Atkins, J. F. Stimulation of stop codon readthrough: frequent presence of an extended 3′ RNA structural element. *Nucleic Acids Res.* **39,** 6679–6691 (2011).
38. Napthine, S., Yek, C., Powell, M. L., Brown, T. D. K. & Brierley, I. Characterization of the stop codon readthrough signal of Colorado tick fever virus segment 9 RNA. *RNA* **18,** 241–252 (2012).
39. Wills, N. M., Gesteland, R. F. & Atkins, J. F. Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci.* **88,** 6991–6995 (1991).
40. Houck-Loomis, B. *et al.* An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature* **480,** 561–564 (2011).
41. Wills, N. M., Gesteland, R. F. & Atkins, J. F. Pseudoknot-dependent read-through of retroviral gag termination codons: importance of sequences in the spacer and loop 2. *EMBO J.* **13,** 4137–4144 (1994).
42. Beier, H. & Grimm, M. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.* **29,** 4767–4782 (2001).
43. Brown, A., Shao, S., Murray, J., Hegde, R. S. & Ramakrishnan, V. Structural basis for stop codon recognition in eukaryotes. *Nature* **524,** 493–496 (2015).
44. Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).
45. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* **2,** 49–58 (2001).
46. Macino, G., Coruzzi, G., Nobrega, F. G., Li, M. & Tzagoloff, A. Use of the UGA terminator as a tryptophan codon in yeast mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 3784–3785 (1979).
47. Böck, A. *et al.* Selenocysteine: the 21st amino acid. *Mol. Microbiol.* **5,** 515–520 (1991).

48. Leinfelder, W., Zehelein, E., Mandrand-Berthelot, M. A. & Böck, A. Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine. *Nature* **331,** 723–725 (1988).
49. Srinivasan, G., James, C. M. & Krzycki, J. A. Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science* **296,** 1459–1462 (2002).
50. Zinoni, F., Heider, J. & Böck, A. Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4660–4664 (1990).
51. Théobald-Dietrich, A., Giegé, R. & Rudinger-Thirion, J. Evidence for the existence in mRNAs of a hairpin element responsible for ribosome dependent pyrrolysine insertion into proteins. *Biochimie* **87,** 813–817 (2005).
52. Zhang, Y., Baranov, P. V., Atkins, J. F. & Gladyshev, V. N. Pyrrolysine and selenocysteine use dissimilar decoding strategies. *J. Biol. Chem.* **280,** 20740–20751 (2005).
53. Stadtman, T. C. Selenocysteine. *Annu. Rev. Biochem.* **65,** 83–100 (1996).
54. Kyriakopoulos, A. & Behne, D. Selenium-containing proteins in mammals and other forms of life. *Rev. Physiol. Biochem. Pharmacol.* **145,** 1–46 (2002).
55. Leibundgut, M., Frick, C., Thanbichler, M., Böck, A. & Ban, N. Selenocysteine tRNA-specific elongation factor SelB is a structural chimaera of elongation and initiation factors. *EMBO J.* **24,** 11–22 (2005).
56. Paleskava, A., Konevega, A. L. & Rodnina, M. V. Thermodynamic and kinetic framework of selenocysteyl-tRNASec recognition by elongation factor SelB. *J. Biol. Chem.* **285,** 3014–3020 (2010).
57. Copeland, P. R., Fletcher, J. E., Carlson, B. A., Hatfield, D. L. & Driscoll, D. M. A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. *EMBO J.* **19,** 306–314 (2000).
58. Palioura, S., Sherrer, R. L., Steitz, T. A., Söll, D. & Simonovic, M. The human SepSecS-tRNASec complex reveals the mechanism of selenocysteine formation. *Science* **325,** 321–325 (2009).

59. Wu, X. Q. & Gross, H. J. The long extra arms of human tRNA((Ser)Sec) and tRNA(Ser) function as major identify elements for serylation in an orientation-dependent, but not sequence-specific manner. *Nucleic Acids Res.* **21,** 5589–5594 (1993).

60. Rudinger, J., Hillenbrandt, R., Sprinzl, M. & Giegé, R. Antideterminants present in minihelix(Sec) hinder its recognition by prokaryotic elongation factor Tu. *EMBO J.* **15,** 650–657 (1996).

61. Baron, C. & Böck, A. The length of the aminoacyl-acceptor stem of the selenocysteine-specific tRNA(Sec) of Escherichia coli is the determinant for binding to elongation factors SELB or Tu. *J. Biol. Chem.* **266,** 20375–20379 (1991).

62. Baranov, P. V. *et al.* RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Res.* **29,** 264–267 (2001).

63. Farabaugh, P. J. Programmed translational frameshifting. *Annu. Rev. Genet.* **30,** 507–528 (1996).

64. Jacks, T. & Varmus, H. E. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. *Science* **230,** 1237–1242 (1985).

65. Jacks, T., Madhani, H. D., Masiarz, F. R. & Varmus, H. E. Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. *Cell* **55,** 447–458 (1988).

66. Jacks, T. *et al.* Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331,** 280–283 (1988).

67. Dulude, D., Berchiche, Y. A., Gendron, K., Brakier-Gingras, L. & Heveker, N. Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. *Virology* **345,** 127–136 (2006).

68. Plant, E. P., Rakauskaite, R., Taylor, D. R. & Dinman, J. D. Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. *J. Virol.* **84,** 4330–4340 (2010).

69. Brierley, I. Ribosomal frameshifting on viral RNAs. *J. Gen. Virol.* **76,** 1885–1892 (1995).

70. Giedroc, D. P. & Cornish, P. V. Frameshifting RNA Pseudoknots: Structure and Mechanism. *Virus Res.* **139,** 193–208 (2009).
71. Plant, E. P. *et al.* The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. *RNA N. Y. N* **9,** 168–174 (2003).
72. Chen, G., Chang, K.-Y., Chou, M.-Y., Bustamante, C. & Tinoco, I. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12706–12711 (2009).
73. Hansen, T. M., Reihani, S. N. S., Oddershede, L. B. & Sørensen, M. A. Correlation between mechanical strength of messenger RNA pseudoknots and ribosomal frameshifting. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 5830–5835 (2007).
74. Ritchie, D. B., Foster, D. A. N. & Woodside, M. T. Programmed −1 frameshifting efficiency correlates with RNA pseudoknot conformational plasticity, not resistance to mechanical unfolding. *Proc. Natl. Acad. Sci.* **109,** 16167–16172 (2012).
75. Chen, J. *et al.* Dynamic pathways of -1 translational frameshifting. *Nature* **512,** 328–332 (2014).
76. Kim, H.-K. *et al.* A frameshifting stimulatory stem loop destabilizes the hybrid state and impedes ribosomal translocation. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 5538–5543 (2014).
77. Yan, S., Wen, J.-D., Bustamante, C. & Tinoco Jr., I. Ribosome Excursions during mRNA Translocation Mediate Broad Branching of Frameshift Pathways. *Cell* **160,** 870–881 (2015).
78. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **244,** 182–188 (1989).
79. Lee, N., Bessho, Y., Wei, K., Szostak, J. W. & Suga, H. Ribozyme-catalyzed tRNA aminoacylation. *Nat. Struct. Biol.* **7,** 28–33 (2000).
80. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the Genetic Code of Escherichia coli. *Science* **292,** 498–500 (2001).

81. Chin, J. W. *et al.* An Expanded Eukaryotic Genetic Code. *Science* **301,** 964–967 (2003).
82. Greiss, S. & Chin, J. W. Expanding the Genetic Code of an Animal. *J. Am. Chem. Soc.* **133,** 14196–14199 (2011).
83. Lang, K. & Chin, J. W. Cellular Incorporation of Unnatural Amino Acids and Bioorthogonal Labeling of Proteins. *Chem. Rev.* **114,** 4764–4806 (2014).
84. Cornish, V. W. *et al.* Site-specific incorporation of biophysical probes into proteins. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 2910–2914 (1994).
85. Cornish, V. W., Hahn, K. M. & Schultz, P. G. Site-Specific Protein Modification Using a Ketone Handle. *J. Am. Chem. Soc.* **118,** 8150–8151 (1996).
86. Summerer, D. *et al.* A genetically encoded fluorescent amino acid. *Proc. Natl. Acad. Sci.* **103,** 9785–9789 (2006).
87. Wang, J., Xie, J. & Schultz, P. G. A genetically encoded fluorescent amino acid. *J. Am. Chem. Soc.* **128,** 8738–8739 (2006).
88. Lee, H. S., Guo, J., Lemke, E. A., Dimla, R. D. & Schultz, P. G. The Genetic Incorporation of a Small, Environmentally Sensitive, Fluorescent Probe into Proteins in S. Cerevisiae. *J. Am. Chem. Soc.* **131,** 12921–12923 (2009).
89. Chatterjee, A., Guo, J., Lee, H. S. & Schultz, P. G. A genetically encoded fluorescent probe in mammalian cells. *J. Am. Chem. Soc.* **135,** 12540–12543 (2013).
90. Fleissner, M. R. *et al.* Site-directed spin labeling of a genetically encoded unnatural amino acid. *Proc. Natl. Acad. Sci.* **106,** 21637–21642 (2009).
91. Kim, C. H., Axup, J. Y. & Schultz, P. G. Protein conjugation with genetically encoded unnatural amino acids. *Curr. Opin. Chem. Biol.* **17,** 412–419 (2013).
92. Lang, K. *et al.* Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. *Nat. Chem.* **4,** 298–304 (2012).
93. Lang, K. *et al.* Genetic Encoding of Bicyclononynes and trans-Cyclooctenes for Site-Specific Protein Labeling in Vitro and in Live Mammalian Cells via Rapid Fluorogenic Diels–Alder Reactions. *J. Am. Chem. Soc.* **134,** 10317–10320 (2012).

94. Uttamapinant, C. *et al.* Genetic Code Expansion Enables Live-Cell and Super-Resolution Imaging of Site-Specifically Labeled Cellular Proteins. *J. Am. Chem. Soc.* **137,** 4602–4605 (2015).

95. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 10092–10097 (1997).

96. Forster, A. C. *et al.* Programming peptidomimetic syntheses by translating genetic codes designed de novo. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 6353–6357 (2003).

97. Anderson, J. C. *et al.* An expanded genetic code with a functional quadruplet codon. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7566–7571 (2004).

98. Neumann, H., Slusarczyk, A. L. & Chin, J. W. De novo generation of mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs. *J. Am. Chem. Soc.* **132,** 2142–2144 (2010).

99. Chatterjee, A., Xiao, H. & Schultz, P. G. Evolution of multiple, mutually orthogonal prolyl-tRNA synthetase/tRNA pairs for unnatural amino acid mutagenesis in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14841–14846 (2012).

100. Rogerson, D. T. *et al.* Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nat. Chem. Biol.* **11,** 496–503 (2015).

101. Sturchler, C., Westhof, E., Carbon, P. & Krol, A. Unique secondary and tertiary structural features of the eucaryotic selenocysteine tRNA(Sec). *Nucleic Acids Res.* **21,** 1073–1079 (1993).

102. Martin, P., Albagli, O., Poggi, M. C., Boulukos, K. E. & Pognonec, P. Development of a new bicistronic retroviral vector with strong IRES activity. *BMC Biotechnol.* **6,** 4 (2006).

103. Schmeing, T. M. & Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461,** 1234–1242 (2009).

104. Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.* **25,** 770–777 (2007).

105. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464,** 441–444 (2010).

106. Orelle, C. *et al.* Protein synthesis by ribosomes with tethered subunits. *Nature* **524,** 119–124 (2015).

107. Fried, S. D., Schmied, W. H., Uttamapinant, C. & Chin, J. W. Ribosome Subunit Stapling for Orthogonal Translation in E. coli. *Angew. Chem. Int. Ed Engl.* **54,** 12791–12794 (2015).

Chapter 2

An In Vitro Selection for Translation Reprogramming: Identification of Eukaryotic Stop Codon Readthrough Signals

---

*The contents of this chapter will be published in:

A.V. Anzalone, V.W. Cornish, et al. "Identification of Eukaryotic Stop Codon Readthrough Signals by *In Vitro* Selection." *In preparation*.

2.0 Chapter Outlook

Translation in eukaryotes can be reprogrammed by cis-acting RNA signals to regulate protein expression and expand the coding capacity of individual genes. From natural examples, it is apparent that a wide variety of RNA features can promote many distinct forms of translation reprogramming. However, this relatively small set of natural cases limits comprehensive evaluation of RNA's translation reprogramming potential, and may not be representative of all reprogramming signals in biology. Therefore, methods for quickly characterizing large numbers of prospective reprogramming signals would advance our understanding of these phenomena, and possibly aid in the identification of naturally occurring reprogramming events. In this chapter, we establish an experimental framework for *in vitro* selection of RNA sequences that reprogram translation, focusing specifically on stop codon readthrough (RT). We first validated our approach, which is based on mRNA display technology, by conducting control and mock selection experiments. Then, to enrich RT promoting RNA sequences *de novo*, we applied the selection to a library containing $>10^{14}$ unique mRNA variants. We performed next generation sequencing of the *in vitro* selection products, then analyzed the data to identify and characterize known as well as novel RT motifs. Lastly, we evaluated the potential for the *in vitro* selected RT motifs to enhance stop codon suppression by nonsense suppressor tRNAs *in vivo*. This work establishes that *in vitro* selections can be used to identify RNA sequences with complex biochemical activities, and furthers our understanding of translation reprogramming signals. Lastly, this selection strategy serves as a platform for the discovery and engineering of new RNA tools for the emerging field of synthetic biology (see **Chapter 3**).

2.1 Introduction

Protein translation, the process by which genetic information is decoded, is invariably performed in all living cells by the ribosome and associated translation components. Despite being confronted by a vast diversity of messenger RNA templates from which a similarly diverse set of polypeptide products is synthesized, the cell's highly evolved protein synthesis machinery is programmed to translate each mRNA according to a set of principal decoding rules that relies on successive interpretation of non-overlapping triplet nucleotide codons[1]. While the overwhelming majority of translation events conform to this standard program with high fidelity[2], there are several known instances in which canonical translation is altered in response to cis-acting elements within the mRNA or nascent polypeptide[3]. These events, generally termed 'translation reprogramming' or 'recoding,' are exploited by nature for various regulatory purposes and serve to expand the synthetic capabilities of the translation apparatus[4].

One example of reprogramming is stop codon readthrough (RT)[5], which can be stimulated by sequence or structural motifs within the mRNA[6]. While most cases of programmed RT in eukaryotic systems have been discovered in plant and mammalian viruses[7], this recoding event has also been observed in various endogenous fungal and animal genes[8–11]. Because RT signals can take various forms and often are structural elements within 3' untranslated regions (UTR)[12–14], they can be difficult to identify based on sequence alone. Moreover, the exact mechanisms by which these reprogramming elements operate are not entirely understood, and in some cases, may be influenced by auxiliary factors[15]. As a result, information about the nature, diversity, and overall frequency of reprogramming signals remains incomplete. A more comprehensive picture

of this phenomenon that better defines RT signals would aid in the identification of naturally occurring stimulatory elements as well as provide additional routes of investigation for unraveling the mechanisms of programmed RT.

Previously, genome-wide analysis has been used to identify potential sites of stop codon readthrough in natural genes. One survey identified underrepresented stop codon contexts in the yeast *saccharomyces cerevisiae* and verified the readthrough activity of these nucleotide sequences[16]. Additionally, phylogenetic analysis tools have been used to predict open reading frames downstream of stop codons on the basis of amino acid conservation[17]. While these analyses provide interesting insight into the prevalence of stop codon readthrough in natural contexts, they depend upon the assumption that either the readthrough events are rare and thus underrepresented, or that the readthrough signals are evolutionarily conserved across organisms. An objective method for evaluating the propensity for RNA sequences to promote reprogramming would be useful for identifying elements or motifs that are not yet appreciated.

One potential avenue for broadly characterizing readthrough signals is by selection or screen of randomized sequences. *De novo* screens for eukaryotic RT signals have previously been conducted in *S. cerevisiae* using a short randomized library of six nucleotides directly 3' of the termination codon[18]. This study concluded that a general stop codon readthrough motif exists with the consensus sequence CAR-NBA (where R represents A or G, N represents any nucleotide, and B represents any nucleotide except C). This study was successful in validating one known readthrough motif, however likely did not find additional motifs due to the constrained library size and limitations to the number of variants that could be screened.

Since many reprogramming elements are composed of RNA structures spanning 50 or more nucleotides, *in vivo* libraries offer only a very limited coverage of sequence space (~11-13 nucleotides) that would likely be required to construct such structures. For example, while yeast serves as a convenient eukaryotic host organism for carrying out genetic screens and selections, the current transformation limit constrains the sequence space to libraries on the order of ~10^6 in *S. cerevisiae*, and screening capabilities often fall short of that figure by several orders of magnitude. Moreover, *in vivo* growth selections are often complicated by unknown or uncontrollable factors that influence cellular fitness and therefore undermine the selection for the desired trait.

In contrast, *in vitro* selections[19–23] offer evaluation of library sizes on the order of 10^{13}-10^{15} starting unique sequences, which can be subjected to selection under highly defined and well-controlled conditions. Since the initial development of SELEX and related technologies, many RNA molecules have been discovered that either bind with high affinity and specificity to diverse ligands[24–26] or catalyze chemical reactions[27]. However, few *in vitro* selections[28] have been reported that identify nucleic acid sequences whose function depends on direct interaction with complex cellular machinery like the translation apparatus.

In this work, we leverage the power of *in vitro* nucleic acid selection to address basic questions regarding the nature of programmed stop codon readthrough signals. Our approach combines the expansive libraries accessible to classical *in vitro* selections and the molecular complexity of a cell lysate. Using mRNA display[22] as a selection tool, we link translation reprogramming events to a selectable outcome—namely, a puromycin ligation reaction. We show that this system can be used to efficiently select for sequences

that promote stop codon readthrough from a library of $>10^{13}$ variants within as few as three cycles of selection. Furthermore, we characterize the selection products with next generation sequencing (NGS) and use this data to identify known as well as previously unreported stop codon readthrough motifs. Finally, we demonstrate in yeast that readthrough signals enhance nonsense suppression by suppressor tRNAs, and thus could potentially be used to boost unnatural amino acid incorporation using amber suppression technologies[29].

2.2 Results

2.2.1 Constructing an *in vitro* selection for translation reprogramming

To select for translation reprogramming elements *in vitro*, we designed a system based on mRNA display that labels transcripts that perform the desired reprogramming function with selectable peptide tags. mRNA display technology[22], which was originally developed as a tool for *in vitro* selections of peptides and proteins, covalently links the phenotype contained within a polypeptide to the amplifiable genotype of the mRNA that encodes it. In this technology, mRNA templates are prepared with the antibiotic puromycin tethered to their 3' terminus and then translated in a cell-free translation system. Puromycin is an aminoacyl-tRNA mimic that can bind to the A-site of the ribosome and react with the peptidyl-tRNA, forming a covalent bond to the c-terminal amino acid residue of the elongating polypeptide. When the ribosome reaches the 3' end of an mRNA-display template, it stalls at an engineered RNA-DNA junction that provides the necessary time and appropriate distance for the tethered puromycin to react (**Fig. 2-1**).

While mRNA-display is conventionally employed to select for functional proteins and peptides, we repurposed it to select for the translation reprogramming activity of an mRNA. Because the puromycin reaction is inefficient at distances greater than 50 nucleotides or fewer than 19 nucleotides from the RNA-DNA junction[30], mRNA-peptide fusion is efficient only when the ribosome translates the entirety of the mRNA transcript. An obstacle to translation, namely an in frame stop codon, should prevent the ribosome from reaching the end of the transcript to form an mRNA-peptide fusion[31]. However, if stop codon readthrough elements exist that promote bypass of the stop codon and translation of the remainder of the mRNA template, then an mRNA-peptide fusion should form (**Fig. 2-1**). The peptide, which is designed to contain affinity purification tags, can then be used to selectively enrich mRNA sequences that formed mRNA-peptide fusions from those that programmed the termination of protein synthesis. These purified mRNAs can then be reverse transcribed and PCR amplified for a subsequent round of selection and analysis (**Fig. 2-2**).

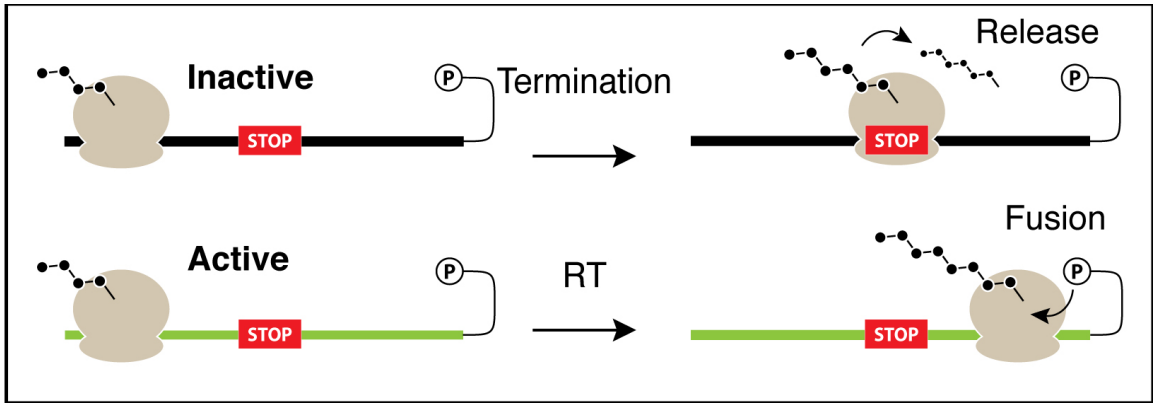


Figure 2-1. Principle of selection for stop codon readthrough using mRNA display. The mRNA is programmed with an in-frame stop codon. If termination occurs, the peptide is released and no mRNA-peptide fusion is formed. If readthrough (RT) occurs, the ribosome can translate the remaining mRNA sequence to form an mRNA-peptide fusion that can be enriched by affinity purification.

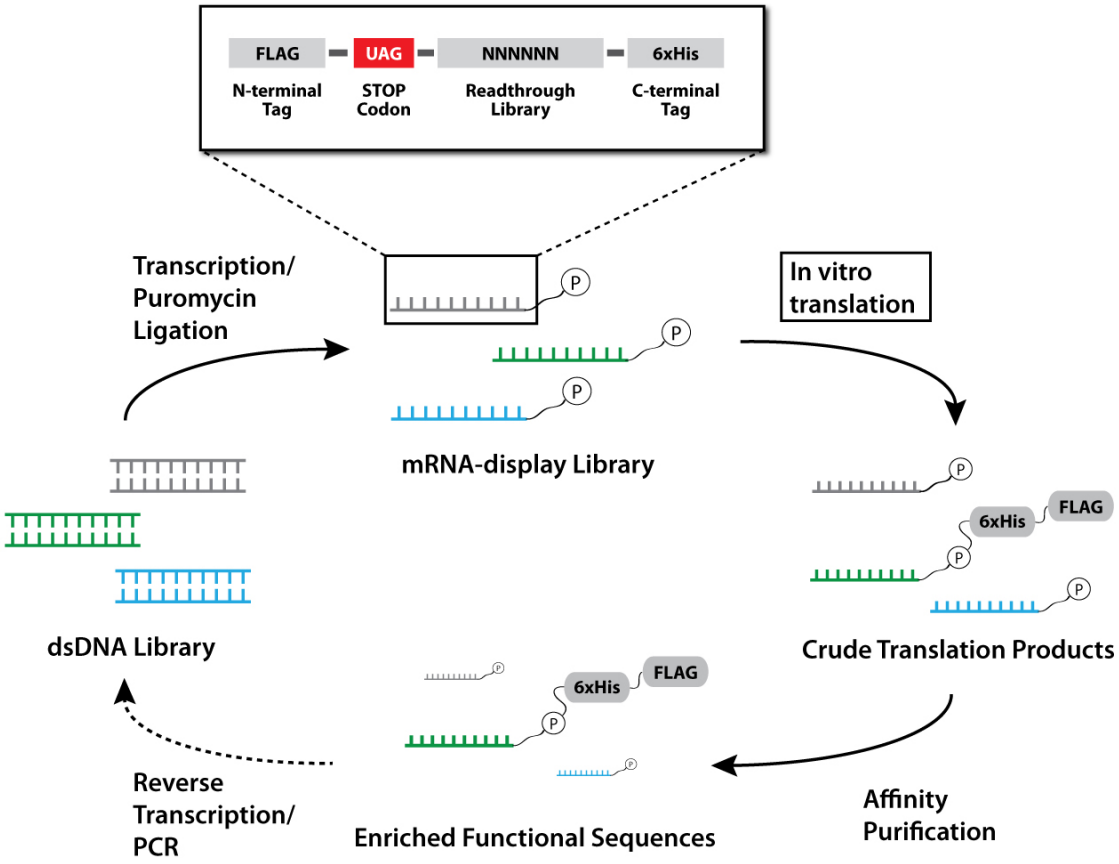


Figure 2-2. mRNA-display selection cycle. First, a dsDNA library is transcribed into RNA and then ligated to the puromycin adaptor. Then, these mRNA-display transcripts are translated *in vitro* in rabbit reticulocyte lysate to form mRNA-peptide fusions. The resulting translation products are affinity purified, reverse transcribed, and PCR amplified for subsequent rounds of selection.

As with previously described mRNA display experiments, our DNA constructs were designed to contain a T7 promoter for transcription of RNA, a TMV translation enhancer element, and both N-terminal and C-terminal affinity purification tags. However, unlike typical mRNA display libraries, our construct also contains a programmed in-frame stop codon that ordinarily triggers translation termination. Thus, the vast majority of mRNA-display templates will be incapable of forming mRNA-

peptide fusions since they likely do not contain RT motifs. N-terminal purification tags serve to exclude sequences that initiate internally at methionine residues downstream of the stop codon. C-terminal purification tags are used to remove sequences that subvert the selection by means of a frameshift mutation, or by premature puromycin fusion reaction.

First, in order to establish that the mRNA-display selection can discriminate against mRNAs that contain stop codons, we generated a control library (10^{14}) that contained *no programmed stop codon* and a stretch of 75 randomized nucleotides in the open reading frame (Control Selection A). Previously, it was shown that mRNA display templates can be pre-selected for open reading frames devoid of stop codons[31], thus purifying transcripts that encode only amino acid residues from those which program termination. The probability that the 75-nucleotide randomized region contains one or more stop codons is ~70%. Thus, after selection, the anticipated result is that the 30% of the sequences within the starting library that lack stop codons in the open reading frame would become enriched. Indeed, after a single round of selection, all 12 sequences sampled were devoid of stop codons in the randomized region (**Table 2-1**), versus 2 of 10 prior to selection (data not shown).

To ascertain the fate of transcripts containing a programmed stop codon but no readthrough motif, we constructed a template containing an in frame UAG stop codon followed by a single randomly generated 66-nucleotide sequence encoding 22 amino acids (Control Selection B). When subjecting this construct to a single round of selection, we unexpectedly found that 5 out of 10 sequences isolated contained mutations to the programmed stop codon and an additional sequence contained a frameshift deletion mutation upstream of the stop codon that placed the UAG out of frame (**Table 2-2**). Only

4 out of 10 sequences were identical to the original construct. Presumably, these mutations arose during *in vitro* transcription of the RNA where the most significant error-prone amplification occurred (vs. plasmid replication). Given that polymerase errors occur with relatively low frequency (T7 RNA polymerase error rate has been estimated at ~$5x10^{-5}$)[32], this result demonstrates that the selection is capable of strongly enriching rare active sequences in a single round of selection. For future rounds of selection, these mutations can be corrected by primer encoded stop codons.

Table 2-1. Control Selection A results. Purifying out nonsense-encoding sequences.

Sample	Sequence	Result
1	P Q D G N S V A W V V G S H A C D W S R E H V G C	No STOP
2	S N F K A I P L C L T L R T R V P D L E R D R V I	No STOP
3	H T A E V T Q W G E S R C R G A L R T F G R C K L	No STOP
4	R F I S R A D Q A S M R S I R I C G A L L G S K Q	No STOP
5	S R C V F G I H R I C G Q G C F R Q L I L V T R V	No STOP
6	A T W Q M I V S W R S I S C R A I Y G R L Q I R G	No STOP
7	S L T G S N S E R I S R G Y R T K A H S G G R I L	No STOP
8	D S S A H F P N I C S G F V L V R I L G V A G T C	No STOP
9	T Y V L F G K R G P M G A Q K N W R V Q G W S E S	No STOP
10	G N V R W F P G D L R R H V T L R V M S Q S N K C	No STOP
11	P R I R C Y A R K V Q I R E L A P R G R H L G C S	No STOP
12	G R I I Y A Y V A D G E I R S V F T T P G D E V S	No STOP

Finally, to demonstrate that reprogramming signals adjacent to stop codons can be enriched from a background of inactive sequences, we performed mock selections (Control Selection C). For the active recoding sequence, we chose the Murine Leukemia Virus (MLV) stimulatory element[12,13,33], a pseudoknot structure that promotes stop codon readthrough at a reported level of 4-5%. Selections were conducted by varying input ratios of the active MLV construct to the inactive control construct. The results, shown in **Table 2-3**, reveal that the active MLV sequence is significantly enriched at an estimated level of ~30-fold after a single round of selection. Therefore, the *in vitro* translation system is capable of responding to reprogramming elements within the mRNA, and the mRNA display selection strategy is capable of enriching active reprogramming elements from a pool composed largely of inactive sequences.

Table 2-2. Control Selection B results. A single sequence containing a stop codon.

Sample	Sequence Description
1	No mutations
2	No mutations
3	No mutations
4	*TAG(Stop) → TGG(Trp)
5	*TAG(Stop) → TAT(Tyr)
6	*ΔA35→Frameshift mutation, no stops
7	*TAG(Stop) → TGG(Trp)
8	No mutations
9	*TAG(Stop) → TGG(Trp)
10	*TAG(Stop) → TGG(Trp)

Table 2-3. Control Selection C results. Mock selection with MLV readthrough signal.

Dilution (MLV : Inactive)	Recovered MLV	Recovered Inactive	MLV Enrichment
1 : 1	5 / 6	1 / 6	5
1 : 10	9 / 12	3 / 12	30
1 : 100	3 / 11	8 / 11	37.5
1 : 1,000	0 / 13	13 / 13	--
Combined 1:10 and 1:100			33.6

2.2.2 *De novo* enrichment of 3' stop codon readthrough motifs

Based on the control and mock selection results described above, we designed a random library aiming to identify readthrough stimulating motifs 3' to the programmed stop codon. Because readthrough stimulating elements can exist in the nucleotides directly adjacent to the stop codon as well as several nucleotides downstream as hairpin or pseudoknot structures, we chose to encode a 75 nucleotide randomized library immediately 3' of a UAG nonsense codon. While this vast theoretical library size (10^{45}) will have exceedingly low coverage from only 10^{14} input molecules, any set of 21 nucleotides from the 75 randomized positions will have >95% coverage. Since many reprogramming stimulatory structures have redundant base-pair possibilities and loops with flexible lengths and compositions, a number of active structural motifs should be accessible without coverage of every nucleotide within that structure.

After a single round of selection on this library, sequences containing mutations to the stop codon were enriched, as was observed in the earlier control selections (Control Selection B). To ensure that these mutant sequences would not take over the library due to their extremely efficient enrichment, primers used for PCR amplification after reverse

transcription were designed to contain the entire 5' segment of the selection construct leading up to and including the TAG stop codon. Therefore, any sequences with stop codon or frameshift mutations either failed to amplify entirely or were corrected, thus preventing false enrichment in subsequent rounds of selection. Additionally, sequences that form aptamers that bind to the anti-FLAG agarose gel and are eluted with FLAG peptide were eliminated by reverse transcription of the RNA (to disrupt RNA structure) and a secondary FLAG purification[34].

After three rounds of selection, sequencing of sampled library members revealed enrichment of a hexanucleotide motif directly adjacent to the UAG stop codon with the consensus sequence CAA-YYA (data not shown). Significantly, this is the same naturally occurring readthrough stimulating sequence used by Tobacco Mosaic Virus (TMV) and generally conforms to the previously identified CAR-NBA motif found in yeast[18]. Enrichment of this sequence motif was highly suggestive that sequences promoting stop codon readthrough were being efficiently selected. Additionally, a small population of sequences were found to contain FLAG peptide epitopes encoded in the library region, suggesting that internal initiation downstream of the stop codon followed by synthesis of a library encoded FLAG epitope is a selectable feature.

To assess the activity of the library members, we developed a fluorescent protein based reporter in the yeast *saccharomyces cerevisiae* that enables a streamlined cloning and assay workflow. Since many eukaryotic reprogramming elements are functional in yeast, sequences enriched from the *in vitro* selection in rabbit reticulocyte lysate are also likely to be active in this assay. The reporter construct contains an N-terminal GFP followed by the stop codon, the library insert and a C-terminal mCherry (**Fig. 2-3**). By

comparison to control constructs that produce only GFP or only GFP-mCherry fusion proteins, reprogramming efficiencies can be determined based on the ratio of the signals from the two fluorescent proteins. Indeed, when selecting samples of yeast colonies, each containing a single library element from the *in vitro* selection, various degrees of stop codon readthrough were observed, with the highest levels reaching 20% (**Fig. 2-3**).

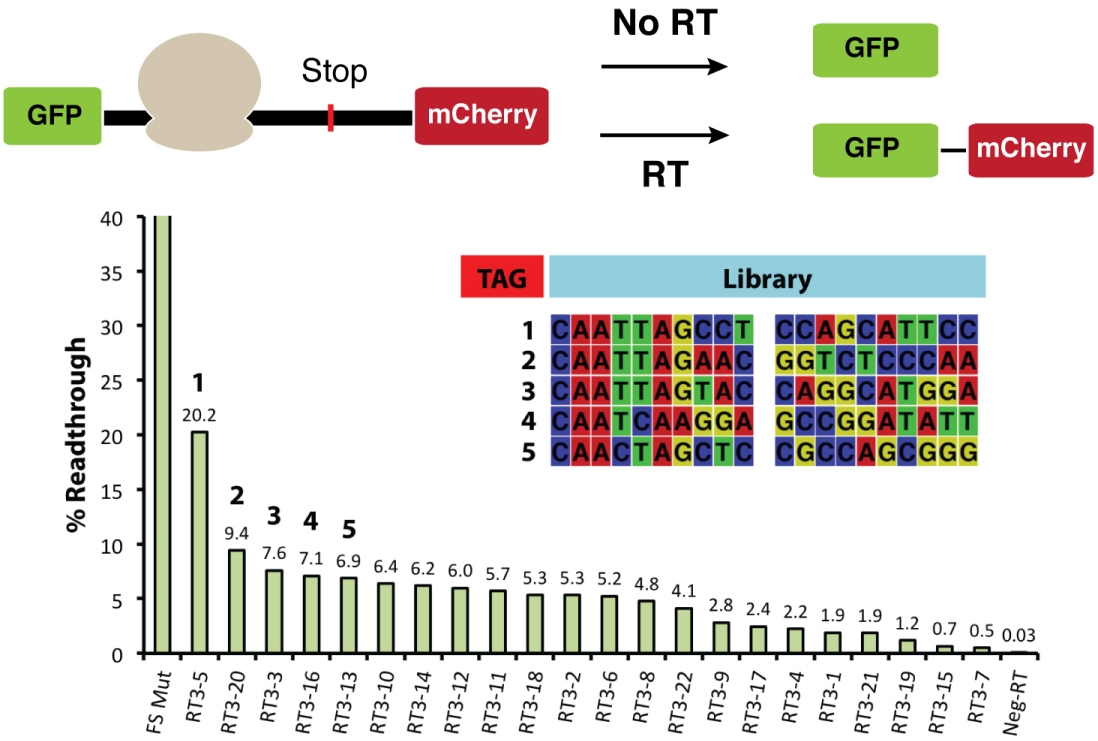


Figure 2-3. Dual-FP reporter assay of RT in *S. cerevisiae*. Selection products were cloned into the dual-FP reporter plasmid and colonies were assayed for readthrough activity by quantification of the fluorescence outputs and comparison to an authentic fusion protein control (FS mut, set to 100%). The CAA-YYA motif is clearly observed in the top sequences, however, given the difference in readthrough efficiencies of these sequences, other factors beyond the CAA-YYA motif apparently contribute to the reprogramming activity of these sequences.

2.2.3 High-throughput sequencing and analysis of RT selection products

To better characterize the landscape of readthrough motifs, we submitted our *in vitro* selection products to high-throughput next generation sequencing (NGS). This

provided ~100 million initial raw sequencing reads. Unique sequences were retrieved from this pool and the total number of reads associated to each sequence was tallied. The sequence list was then filtered to remove members that encoded library FLAG epitopes or internal stop codons, reducing the set to ~11 million unique sequences. Additionally, because the remaining sequences were significantly composed of variants that were simple nucleotide substitutions or frameshift mutations away from a parent sequence, all duplicates with a hamming distance of < 10 were removed from the top-ranked 1 million unique sequences. This reduced the pool of sequences to a final set containing ~42,000 truly unique members, which were used for the analysis. These sequences account for ~43 million of the initial ~100 million sequencing reads. Within this group, 50% of the read mass (21.5 million) is contained within the top ~10,000 sequences (**Fig. 2-4**).

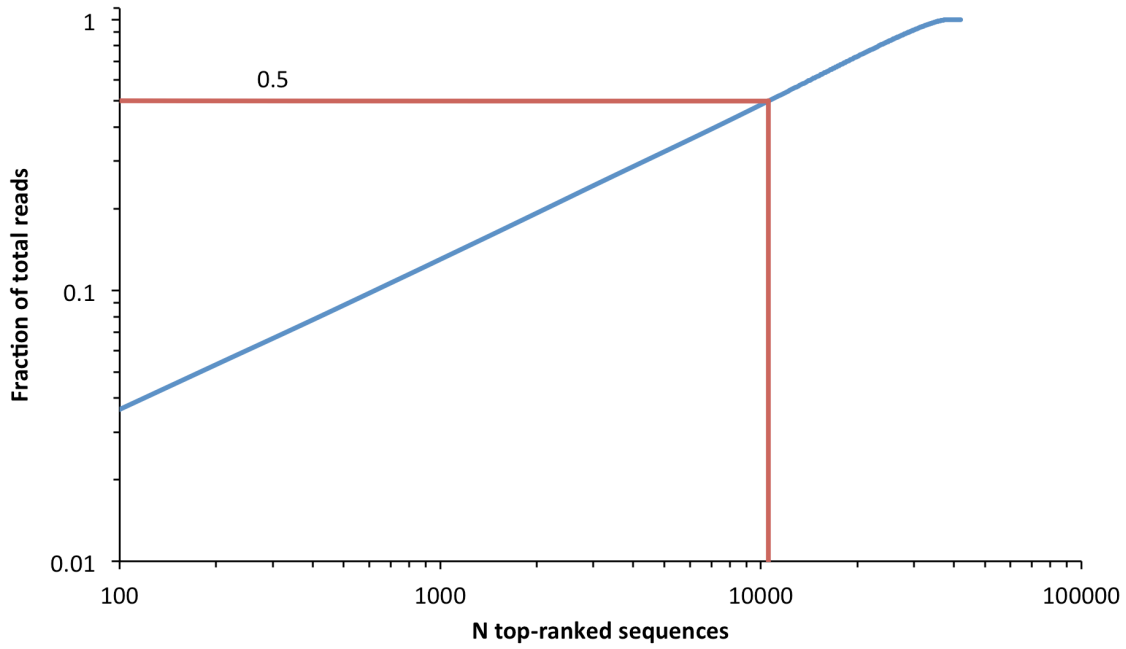


Figure 2-4. Distribution of NGS sequence mass. Composition of the library mass (reads) contributed by the top N sequences is plotted (blue). 50% of the total reads are contributed by the top ~10,000 unique sequences (highlighted by red lines).

We first analyzed the prevalence of each of the four nucleotides in the position directly adjacent to the amber stop codon (hereafter referred to as position +1). Within the 500 top-ranked unique sequences (ranked by read count), almost every member contained a C at the +1 position (497 out of 500), and nearly 99% of the read mass contained within this group was contributed by these sequences. This is consistent with previous reports that a cytidine nucleotide adjacent to the stop codon establishes an inefficient termination context[35]. In the total set of ~42,000 sequences, 76% contain a C at the +1 position, accounting for 85.5% of the sequencing read mass. The next most prevalent nucleotide at the +1 position was G, which represents 12% of total sequences and 8% of the read mass for the entire sequence pool (**Table 2-4**).

Table 2-4. Prevalence of nucleotides 3' adjacent to stop codon (+1).

Nucleotide	**Top 500 sequences**		**Top 42,000 sequences**	
	Number of sequences	*% of read mass*	*% of total sequences*	*% of total read mass*
C	497	98.8	76.0	85.5
G	2	1.0	12.0	8.0
A	0	0	6.9	4.0
U	1	0.1	5.1	2.5

Next, we evaluated the prevalence of the three nucleotides adjacent to the stop codon, occupying positions +1 to +3 (**Table 2-5**). For this triplet of nucleotides, there is a strong enrichment for the CAA codon, particularly in the top 500 sequences where it is found 442 times and accounts for roughly 91% of this group's read mass. Three other codons were also significantly enriched at this position: CAG, CCA, and CTA.

Table 2-5. Prevalence of first (+1 to +3) and second (+4 to +6) codons.

Position	Sequence	Top 500 sequences		Total library	
		Number of sequences	*% of read mass*	*% of total sequences*	*% of total read mass*
Codon 1 (+1 to +3)	*CAA*	442	90.8	41.1	52.6
	CAG	22	3.4	11.4	11.8
	CCA	29	4.2	6.1	6.8
	CTA	1	0.1	5.0	4.4
Codon 2 (+4 to +6)	*TTA*	159	36.0	17.1	20.2
	TCA	145	28.3	12.2	16.3
	CTA	78	15.8	8.3	10.6
	CCA	52	8.6	5.7	7.3
	CGA	47	7.7	5.5	6.9

Likewise, the next downstream codon (positions +4 to +6) was enriched for specific triplets. Significantly, the five most abundant codons at this position conform to a consensus YBA motif (where Y is either T or C, B is not A; and, the TGA stop codon is excluded), with TTA being the most abundant (**Table 2-5**). Overall, the enrichment profile of these first two codons is suggestive of a broader CAR-YBA motif. To determine whether a true association exists between these codon pairs, we performed chi-squared and Fisher's exact statistical tests. Both of these methods were used to evaluate the 2x2 contingency table containing the observed frequencies of the codons under consideration pairing with one another or any other codon. The outputs of the tests are various statistics that reflect the significance of the association (χ^2 and p-value) and degree of correlation (ϕ and odds ratio).

The statistics for a select group of sequence pairs are summarized in **Table 2-6**. All combinations of CAA with YBA codons showed large positive ϕ coefficients and χ^2 values (strongest for TTA and TCA), suggesting the presence of a consensus CAA-YBA motif. All together, the CAA-YBA motif is observed at the beginning of 34% of the total sequences in the analysis pool. Of note, CAG in the first codon position was less strongly associated with YBA codons (consider ϕ coefficients, **Table 2-6**), suggesting that CAG may be part of distinct motifs outside of an overall CAR-YBA motif (where R is A or G). However, with the exception of CAG-TCA, positive association of CAG with YBA codons was still observed at a statistically significant level. Additionally, we observed large test statistics for association between CCA at codon 1 and CBA triplet at codon 2, with little correlation observed between CCA and TTA.

Interestingly, when codon 1 was CTA, we observed large negative test statistics for all YBA codons. In fact, the CTA-TTA pair represented one of the largest negative associations for all codon pairs. Instead, the CTA codon was highly associated with G at the +4 position (**Table 2-6**), with an additional slight preference for G at the +5 position (not shown). Notably, recent studies have revealed the CTAG motif in four human genes that display stop codon readthrough[8]. Additionally, a previously described stop codon readthrough signal from Sindbis virus also possess the CTAG motif adjacent to the stop codon, though readthrough was attributed predominantly the to single cytidine nucleotide[35].

Table 2-6. Chi-squared and Fisher's exact test results for statistical association between select sequence pairs.

Codon 1 (+1 to +3)	Codon 2 (+4 to +6)	ϕ	χ^2	Stand. residual	Odds ratio	p-value[a]
CAA	*TTA*	0.312	3891	42.8	5.7	< 0.0001
	TCA	0.276	3042	39.0	6.1	< 0.0001
	CTA	0.127	646	18.4	2.5	< 0.0001
	CCA	0.084	284	12.4	2.0	< 0.0001
	CGA	0.073	214	10.8	1.9	< 0.0001
CAG	*TTA*	0.033	43	5.6	1.3	< 0.0001
	TCA	-0.051	106	-9	0.6	< 0.0001
	CTA	0.045	79	8.0	1.5	< 0.0001
	CCA	0.053	114	9.8	1.8	< 0.0001
	CGA	0.040	64	7.4	1.6	< 0.0001
CCA	*TTA*	-0.014	7.6	-2.4	0.9	0.005
	TCA	0.034	47	6.2	1.5	< 0.0001
	CTA	0.128	651	23.7	3.5	< 0.0001
	CCA	0.071	206	13.5	2.5	< 0.0001
	CGA	0.103	423	19.4	3.3	< 0.0001
CTA	*TTA*	-0.103	424	-18.2	0.03	< 0.0001
	TCA	-0.078	240	-14.1	0.1	< 0.0001
	CTA	-0.067	182	-12.6	0.05	< 0.0001
	CCA	-0.051	106	-9.8	0.1	< 0.0001
	CGA	-0.067	182	-12.6	0.05	< 0.0001
Codon 1 (+1 to +3)	**Position +4**	**ϕ**	**χ^2**	**Stand. residual**	**Odds ratio**	**p-value[a]**
CTA	*G*	0.233	2180	40.9	6.9	< 0.0001
	A	0.011	4.4	2.0	1.15	0.03
	T	-0.115	526	40.9	6.9	< 0.0001
	C	-0.08	295	-14.2	0.3	< 0.0001

[a]This is the non-adjusted p-value output of the Fisher's exact test. Negative correlations are shaded gray.

In the third codon position, CAG was the most common triplet sequence, nearly twice as prevalent as the next most abundant triplet. CAG at codon 3 was also positively associated with the CAA-YBA motif ($\phi = 0.096$, $\chi^2 = 366$). Overall, the sequence CAA-TTA-CAG is observed 532 times, and accounts for almost 3 times as many reads as the next best codons following CAA-TTA (**Fig. 2-5**). Significantly, the CAA-TTA-CAG is precisely the wild-type TMV readthrough sequence. While the third codon is not commonly cited for its contribution to readthrough efficiency, initial characterization of the TMV sequence demonstrated a noticeable decrease in readthrough when this codon was omitted[36]. By comparison, the triplet GGA is highly underrepresented following CAA-TTA, especially when compared to its abundance in other library regions. We evaluated this effect using the dual-FP reporter in yeast. Mutating a selection product from a sequence beginning CAA-TTA-CAG to CAA-TTA-GGA resulted in a 2-fold decrease in readthrough efficiency (9.1% down to 4.4%), providing strong evidence that the codon following CAA-TTA also plays a role in reprogramming.

Previously unreported motifs were uncovered by contingency analysis, which was particularly revealing when the statistical tests were performed beyond the arbitrary codon framework. One especially strong association was found between CGC and CAG triplets spanning position +1 to +6 ($\phi = 0.231$, $\chi^2 = 2135$). This association can be further extended to a CGC-CAG-R consensus motif if the +7 nucleotide is considered in the analysis ($\phi = 0.051$, $\chi^2 = 111$). A substantially longer motif was identified when analysis of the CAG-ACT codon pair ($\phi = 0.060$, $\chi^2 = 147$) was extended to downstream nucleotide quartets (+7 to +10), revealing a motif spanning +1 to +10 with the consensus sequence CAG-ACT-YMM-G ($\phi = 0.141$, $\chi^2 = 843$).

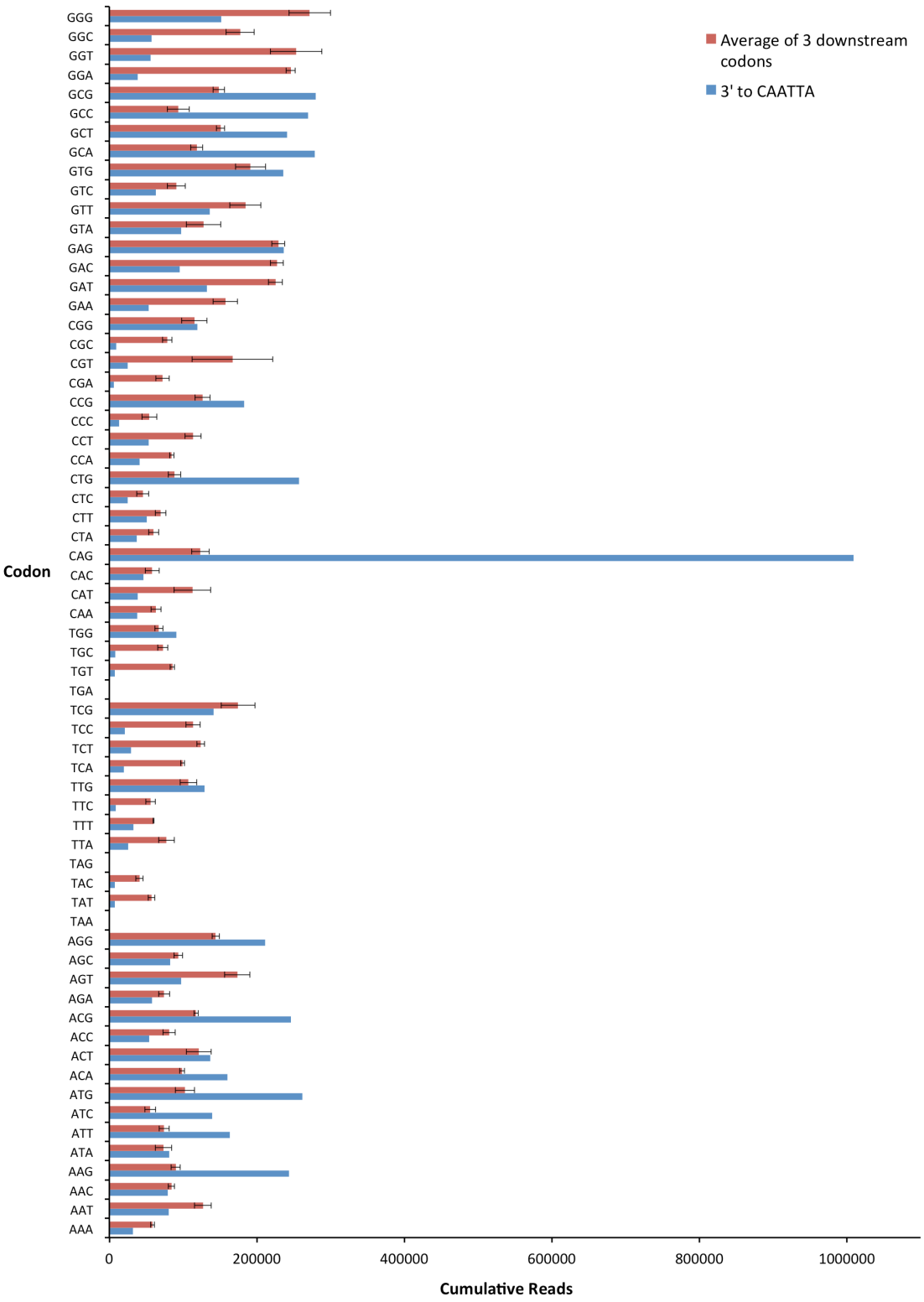


Figure 2-5. Codons adjacent to CAATTA. Cumulative reads for all codons directly 3'-adjacent to CAA-TTA (blue) or significantly downstream (codons 10, 11 and 12) of the stop codon (red).

2.2.4 Enhanced amber suppression with 3' RT sequences

Unnatural amino acid mutagenesis has tremendous potential as a tool for basic science and biotechnology[29]. While numerous analogs of unnatural amino acids have been successfully incorporated *in vivo* by genetic encoding with evolved aaRS enzymes[37], the low yield of incorporation remains a major barrier to widespread adoption of the technology in eukaryotes. One of the primary competing processes for unnatural amino acid incorporation at stop codons is peptide hydrolysis by release factors. Release factor knockouts in *E. coli* were shown to enhance stop codon suppression[38], however this also dramatically increased the off-target incorporation of UAAs at natural stop codons in other cellular mRNAs. Additionally, because eukaryotes have just a single dedicated release factor for recognition of all stop codons, knockout in this setting is not viable. Therefore, we envisioned applying the *in vitro* discovered readthrough signals to improve incorporation efficiency in a way that is specific to the target mRNA.

We sought to determine if the library readthrough signals could enhance suppression efficiency in yeast using an orthogonal *E. coli* tyrosine amber suppressor tRNA/aaRS pair. This assay was performed using the dual fluorescent protein reporter containing the library reprogramming signals. The results, shown in **Figure 2-6**, demonstrate significant enhancement of suppression that is correlated to, but not a direct function of, readthrough in the absence of suppressor tRNA. In the case of low basal readthrough, tRNA-mediated suppression was observed at levels that ranged between 5% and 15%. Impressively, when examining the reprogramming signals that promote higher levels of readthrough, we found that readthrough levels in the presence of suppressor tRNA reached as high as > 60%. This strongly supports the hypothesis that amber

suppression competes with RF hydrolysis, and that inhibition of release enhances suppression efficiency by suppressor tRNAs.

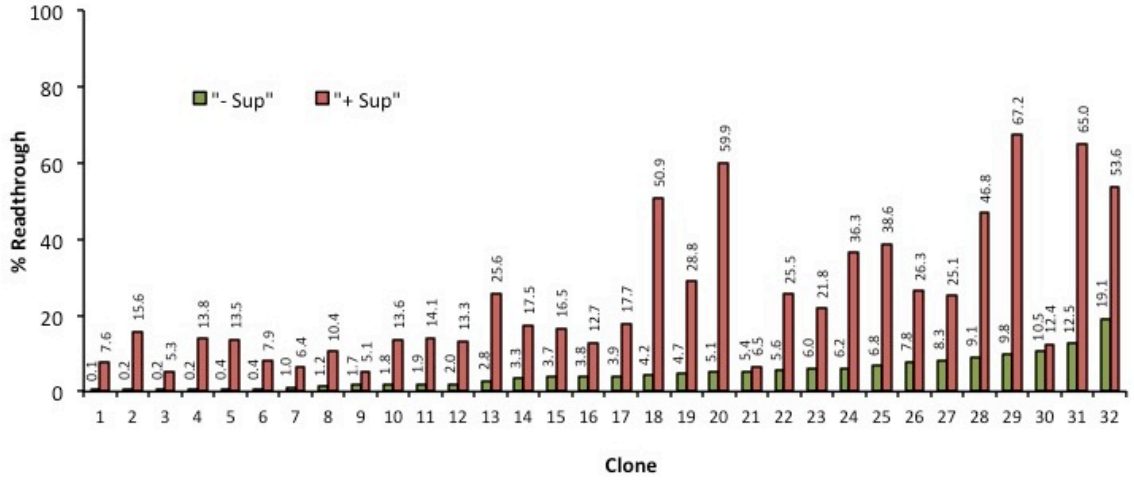


Figure 2-6. Enhanced amber suppression with 3' RT motifs. Readthrough efficiency of sampled reprogramming elements in the absence (green) and presence (red) of tyrosine amber suppressor tRNA.

2.3 Discussion

RNA plays a central role in reprogramming the protein synthetic machinery. While naturally occurring signals have provided examples of RNA motifs and structures that can promote reprogramming, the diversity of these features makes exhaustive analysis of potential reprogramming sequences a major difficulty. To address these challenges, we developed a strategy to *in vitro* select for RNA elements that reprogram the translation machinery using mRNA display, enabling the search of vast libraries of randomized RNA sequence in excess of 10^{13} starting members. Significantly, this *in vitro* RNA selection differs from previous SELEX-type experiments in that it selects for an RNA function that requires direct collaboration with complex cellular machinery.

Because the selection is conducted in a cell lysate, the large library sizes that are accessible to typical RNA selections are maintained, while gaining molecular complexity on par with an in *vivo* system.

The control experiments performed here validate the hypothesis that full translation of an mRNA transcript can be used as a general selectable outcome with mRNA display. The principle behind this selection is analogous to the use of downstream reporter genes in an *in vivo* setting. Importantly, control experiments also demonstrated that mRNA display could be used to select for authentic translation reprogramming elements and significantly enrich these sequences from a background of predominantly inactive sequences. Overall, these results establish mRNA display as an effective tool for the *in vitro* selection of translation reprogramming elements.

We then applied the mRNA display selection to search a large library of RNA sequences (10^{14} variants) for stop codon readthrough signals. Significantly, after just three rounds of *in vitro* selection, this yielded highly active readthrough elements as determined by an *in vivo* dual-FP assay in *S. cerevisiae*. The products of the selection were submitted to high-throughput sequencing in order to identify and characterize various stop codon readthrough signals. An overwhelming prevalence of a hexanucleotide motif with the consensus sequence CAA-YBA was observed. Importantly, this is a well-established readthrough signal, typified by the TMV readthrough sequence CAA-TTA. By analyzing the group of sequences containing the CAATTA motif, we also discovered that the codon directly 3' to this sequence also influences readthrough efficiency. Aside from this well-established motif, we also found significant enrichment for a CTAG motif directly adjacent to the stop codon, which has

recently been found to stimulate stop codon readthrough in four human genes and several fungal genes. Lastly, analysis of the NGS data was used to identify previously unreported motifs, namely CGC-CAG-R and CAG-ACT-YMMG. Further analysis of this sequence pool could identify structural elements, though these are likely to be far more diverse at the sequence level.

Beyond characterization of readthrough elements, this selection potentially provides useful parts for emerging biotechnologies. We demonstrated that stop codon readthrough signals discovered from *in vitro* selection are capable of significantly enhancing nonsense suppression with amber suppressor tRNAs. This has potential applications to the field of unnatural amino acid mutagenesis, which utilizes suppressor tRNAs to direct the incorporation of non-standard amino acids into proteins. Overall, this selection strategy provides a framework for engineering new devices for synthetic biology (see **Chapter 3**).

2.4 Experimental Methods

General materials and methods. Standard methods for molecular biology in *Saccharomyces cerevisiae* and *Escherichia coli* were used[39]. Oligonucleotides used in this study are listed in **Table 2-7** (chapter **2.5**). DNA sequences of the selection constructs are outlined in **Table 2-8** (chapter **2.5**). Unless otherwise specified, PCR reactions were carried out with Vent Polymerase (NEB) in 1X Thermopol buffer. All steps involving RNA were performed with RNase-free materials and RNase-free ddH_2O (Millipore) at 4 ˚C unless otherwise specified. RNA was kept at -80 ˚C for long-term storage.

PCR assembly of selection constructs. Control constructs and libraries were assembled by PCR from chemically synthesized oligonucleotides. All dsDNA constructs contain a 5' T7 promoter, a TMV translation enhancer sequence, ATG start codon, N-terminal FLAG peptide, reprogramming segment, and C-terminal hexahistidine tag. The library for Control Selection A was constructed by PCR amplification of the CT-Lib oligonucleotide with CT-For and CT-Rev in a 1-mL reaction. The Inactive-RT construct used in Control Selections B and C was assembled by polymerase extension of NegRT-Inner-for with NegRT-Inner-rev, followed by PCR amplification with NegRT-For and NegRT-Rev. The Murine Leukemia Virus (MLV) readthrough construct used in Control Selection C was prepared by PCR fusion of two dsDNA segments: one derived from polymerase extension of MLV-PK-1 with MLV-PK-2, and another derived from polymerase extension of AVA110-MLV-PK with MLV-PK-4. The fused product was amplified with MLV-PK-1 and MLV-PK-4. The 75-nucleotide randomized readthrough library was constructed by PCR amplification of the library oligonucleotide RT-Lib with RT-For and RT-Rev in a 1-mL reaction.

***In vitro* transcription of RNA templates.** RNA was *in vitro* transcribed from dsDNA templates (250 nM) using T7 RNA polymerase. Transcription reactions were performed with T7 RNA polymerase (NEB, 6uL/100uL reaction) in T7 buffer (50 mM Tris, pH 8, 20 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, 0.01% Triton-X100) supplemented with NTPs (5 mM final of each NTP) at 37 ˚C. After 2-3 hours, if the transcription reaction produced a white precipitate, supplemental $MgCl_2$ (5 mM) was added to the reaction. Total reaction times ranged between 4 and 5 hours. Transcription reactions were terminated by the addition of 0.1 volumes of 0.5 M EDTA, pH 8. For the

first round transcription of the CT and RT libraries, reactions were performed at the 1-mL scale to provide adequate library coverage and sufficient quantities of RNA. RNA was purified by denaturing polyacrylamide gel electrophoresis, typically 7.5% Urea-PAGE. Following electrophoretic separation (250V, 50 minutes, 1X TBE), RNA bands were visualized by UV-shadowing and excised from the gel. The RNA was eluted from the gel slab by electro-elution (15 minutes, 225V, 1X TAE) in dialysis tubing (10,000 MWC). The eluted RNA was buffer exchanged into 0.3 M NaOAc (pH 5.2) by spin filtration (Amicon 10,000 MWC), then precipitated by the addition of 2.3 volumes of 100% ethanol and storage overnight at -80 ˚C. The precipitated RNA was pelleted by centrifugation (4 ˚C, 14,000 rpm) and air-dried. The pellet was then resuspended in an appropriate volume of RNase-free ddH_2O (to roughly 100 uM final) and the concentration measured by UV-absorbance. If necessary, RNA solutions were further diluted to the desired concentration.

Preparation of mRNA-display templates. mRNA-display templates were prepared by splint ligation of the *in vitro* transcribed RNA to the puromycin adaptor (phospho-dA_{27}-CC-puromycin) using T4 RNA ligase (NEB, 2,000,000 U/mL). The RNA (10 uM final), phospho-dA_{27}-dCdC-puromycin (10 uM final), splint oligonucleotide AVA95 (15 uM final), splint oligonucleotide AVA96 (5 uM final), and RNase-free ddH_2O were combined and heated to 95 ˚C for 5 minutes. The mixture was then treated with the appropriate volume of 10X T4 ligase buffer (1X final) and quickly transferred directly to ice and cooled for 5 minutes. The reaction was then allowed to warm to room temperature for 5 minutes and treated with T4 DNA ligase (40,000 U/mL final). The reaction was left at room temperature for 2 hours, then treated with 0.5 M EDTA, pH 8 to

a final concentration of 10 mM. The ligation reactions were separated using 7.5% Urea-PAGE gels and the slower migrating ligation product (upper band) was excised. RNA-puromycin products were isolated similarly to the RNA transcripts from the previous step by elution from the PAGE gel and precipitation.

In vitro **translation and mRNA-peptide fusion purification**. General considerations: Translations are conducted in 40% rabbit reticulocyte lysate (nuclease treated, Promega) with mRNA-display template concentrations ranging from 200 to 800 nM. Rabbit reticulocyte lysate should be thawed just prior to use and distributed in aliquots after the first thawing to avoid multiple freeze-thaw cycles. Rabbit reticulocyte lysate should not be centrifuged. Rabbit reticulocyte lysate should always be the last component added to the reaction mixture. The first round translation reaction for the readthrough library was performed at a 1-mL scale. The following translation and purification protocol serves as a representative selection round performed at the 100-uL scale. FLAG purifications[34] and Ni-NTA purifications[30] were performed essentially as described.

In vitro translation. Components are combined at 4 ˚C in the order as listed: mRNA-display template (10 mM, 2 uL), amino acid mix (50X, 2 uL), KCl (2.5 M, 4 uL), $Mg(OAc)_2$ (25 mM, 2 uL), RNase-free ddH_2O (50 uL), and freshly thawed rabbit reticulocyte lysate (2.5X, 40 uL). The reaction is mixed well and incubated at 30 ºC for 1 hour. After incubation, 38 uL of puromycin salt mix (31.2 uL of 2.5M KCl + 6.8 uL 1M $MgCl_2$) is added to the reaction, which is then left to incubate at room temperature for 15 minutes.

FLAG purification #1. The translation reaction is then diluted 10-fold by adding 9 volumes (1.24 mL) of 1X FLAG binding buffer, then incubated at 4 ºC with rotation for 1.5 hours with 30 uL of M2-anti-FLAG resin (prewashed 2x with 0.75 mL FLAG clean buffer and 3x with 0.75 mL 1X FLAG binding buffer). The resin is then gently pelleted (1,000 g) and the supernatant removed. The resin is washed with 0.75 mL FLAG binding buffer, gently pelleted, and the supernatant removed. This is repeated with 5 additional FLAG binding buffer washes. mRNA-peptide fusions are then eluted 2x by incubating the resin at 4 ºC with rotation for 20 minutes with 100 uL of FLAG peptide in FLAG binding buffer (200 ug/mL FLAG peptide).

Ni-NTA purification. The FLAG elutions are combined and mixed with 1 volume of 2X Ni-NTA binding buffer, then incubated at 4 ˚C with rotation for 1 hour with 100 uL of Ni-NTA resin (prewashed 3x with RNase-free ddH_2O). The resin was then gently pelleted (1,000 g) and the supernatant removed. The resin is then washed with the following solutions, each 0.75 mL: (i) Ni-NTA Wash Buffer-1 (2x); (ii) a 4:1 mix of Ni-NTA Wash Buffer-1 to Ni-NTA Wash Buffer-2; (iii) a 3:2 mix of Ni-NTA Wash Buffer-1 to Ni-NTA Wash Buffer-2; (iv) a 2:3 mix of Ni-NTA Wash Buffer-1 to Ni-NTA Wash Buffer-2; (v) a 1:4 mix of Ni-NTA Wash Buffer-1 to Ni-NTA Wash Buffer-2; (vi) Ni-NTA Wash Buffer-2; and (vii) a 19:1 mix of Ni-NTA Wash Buffer 2 to Ni-NTA Elution Buffer. mRNA-peptide fusion are then eluted from the resin 2x with 100 uL portions of Elution buffer by incubation at 4 ˚C with rotation for 30 minutes. The combined elutions are spin-concentrated to a volume of ~30 mL (0.5-mL 30,000 MWC amicon) and 1 uL of the concentrated elution is saved for the non-reverse transcribed PCR control reaction.

Reverse transcription of cDNA from purified mRNA-peptide fusions. The reverse transcription reaction is prepared by adding the following components in the order listed: concentrated Ni-NTA elution (30 uL), reverse transcription primer AVA95 (100 uM, 3 uL), dNTP mix (10 mM, 15 uL), RNase-free ddH_2O (160 uL). These components are mixed and incubated for 5 minutes at room temperature, then treated with: 5X first-strand buffer (5X, 60 uL), DTT (0.1 M, 30 uL), RNasin (1 uL), and Superscript II Reverse Transcriptase (1 uL). The reaction is mixed well and incubated at 42 ºC for 50 minutes.

FLAG purification #2. An optional second FLAG purification can be performed to remove undesired RNA aptamers from the selection pool. After incubation, the reverse transcription reaction is spin concentrated and buffer exchanged (0.5-mL 30,000 MWC amicon) into 0.5 mL FLAG binding buffer. The FLAG purification is performed essentially as described for FLAG purification #1. The combined elutions are spin concentrated to ~30 uL for PCR.

PCR amplification of cDNA. Analytical PCR reactions are prepared for controls with no RT reaction (ddH_2O as template) and pre-reverse transcription mRNA-peptide fusions (concentrated FLAG elution), in addition to the post-reverse transcription product. The following components are combined in the order listed for a 50-uL scale reaction: PCR template (0.5 uL), forward primer (100 uM, 0.5 uL), reverse primer (100 uM, 0.5 uL), dNTP mix (10 mM, 1 uL), 5X GoTaq Buffer (5X, 10 uL), ddH_2O (37 uL), GoTaq Polymerase (0.5 uL). If analytical reactions produce the desired results, no PCR product for ddH_2O as template or pre-reverse transcription mRNA-peptide fusions as template, a PCR is performed at a 0.5 mL scale with the post-reverse transcription

product as a template. This PCR product serves as the dsDNA input for the subsequent round of selection, and as template for cloning into reporters or sequencing vectors.

Dual-fluorescent protein reporter for determining RT activity in *S. cerevisiae*. The reporter plasmid was constructed by inserting the yEGFP open reading frame (amplified with primer pair AVA111·AVA112) between the GPD promoter and the open reading frame of a yeast-optimized mCherry[40] flanked by a GPD terminator in a pRS425 backbone (high copy, *LEU2* marker). NheI and AatII restriction sites were encoded between the GFP and mCherry open reading frames for replacement of the intervening sequence with a RT insert. To assay library readthrough activity, dsDNA products from the *in vitro* selection were PCR amplified in two steps (primer pairs AVA136·AVA118 and AVA119·AVA118) to add homology to the NheI/AatII digested dual-FP reporter plasmid. The library was cloned by *in vivo* gap repair into yeast strain Fy251 using high efficiency transformation as previously described[41] and plated on selective medium (synthetic complete agar with leucine dropout containing 2% dextrose; SC-Agar-(Gluc) L-). Individual colonies were isolated and grown in liquid SC-(Gluc) L- media to mid-log phase and the readthrough efficiency was determined by comparing quantified GFP and mCherry fluorescence signals to a fusion protein control (calibrated to 100% readthrough).

Next-generation sequencing. NGS was performed using the Illumina HiSeq platform (Columbia Genome Center). DNA was prepared for sequencing by PCR in three separate batches containing either 0, 1, or 2 variable nucleotides at the start site of the HiSeq read to avoid identical base calls during the initialization phase of sequencing. The third round *in vitro* selection products were PCR amplified with primer pair

AVA314·AVA320, AVA315·AVA320 or AVA316·AVA320. All three PCR products were then mixed and amplified with primer pair AVA319·AVA321 to add full HiSeq adaptor sequences. After sequencing, the raw fastq file was filtered to remove FLAG-epitope containing sequences or sequences that encode an in-frame stop codon. The copy number of each remaining unique sequence was computed. From this set of sequences, the 1 million top-ranked (by abundance) members were evaluated for their true uniqueness (defining truly unique as representing an original starting library member derived from chemical synthesis). Sequences that were single nucleotide substitutions, or frameshift mutations, away from a higher ranked parent sequence were removed by setting a hamming distance threshold. This reduced the set of sequences to ~42,000, and these sequences were used for the subsequent analysis.

Analysis of NGS sequencing data. NGS sequencing results were evaluated using simple counting algorithms, Fisher's exact tests, and chi-squared tests implemented in customized python code. For calculating the association between two variables, for instance the identity of codon 1 having the sequence CAA and the identity of codon 2 having the sequence TTA, a 2 x 2 contingency table was constructed containing the following elements: the occurrence (number of unique sequences) of CAA and TTA; the occurrence of CAA and *not*TTA; the occurrence of *not*CAA and TTA; and, the occurrence of *not*CAA and *not*TTA. Expected frequencies that would result for null association were then calculated based on this contingency table. Fisher's exact and chi-squared tests were used to determine the statistical significance of the association (χ^2 and p-value) and degree of correlation (ϕ and odds ratio). For any pair of sequence segments (for instance +1 to +3 analyzed for association with +4 to +6), the aforementioned

statistical tests were performed over the set of all pairs of sequences contained within the data set.

2.5 DNA sequences

Table 2-7. Sequences of oligonucleotides used in this study.

Name	**Sequence (5' to 3')**
CT-Lib	G ACT ACA AGG ACG ACG ACG ACA AGT ACN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNG GCA GCG GCC ATC ATC ACC ATC
CT-For	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG ACG ACG ACA AGT AC
CT-Rev	ATA GCC AGA TCC AGA CAT TCC CAT AGA ACC GCC GTG GTG ATG GTG ATG ATG GCC GCT GCC
NegRT-Inner-for	CGA CGA CAA GAC CCT AGA TGA CTA GCA CCG AGA ACC CGG CGC CAC GCA
NegRT-Inner-rev	GAT GGT GAT GAT GGC CGC TGC CAT ACG TTC CCT TTA ATT GCC TGC CGG
NegRT-For	CAC CGA GAA CCC GGC GCC ACG CAA TGG AAC GTC CTT AAC TCC GGC AGG CAA TTA AAG GGA ACG TAT
NegRT-Rev	ATA CGT TCC CTT TAA TTG CCT GCC GGA GTT AAG GAC GTT CCA TTG CGT GGC GCC GGG TTC TCG GTG
MLV-PK-1	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG ACG ACG ACA AGA CCC TAG ATG ACT AG
MLV-PK-2	CTG GGT TCA GGG GGG GGC TCC TGA CCC TGA CCT CCC TAG TCA TCT AGG GTC TTG TCG TCG
AVA110-MLV-PK	GGA GCC CCC CCC TGA ACC CAG GAT AAC CCT CAA AGT CGG GGG GCA ACC CGT CGG CAG CGG CCA TCA TCA CCA TC
MLV-PK-4	ATA GCC AGA TCC AGA CAT TCC CAT AGA ACC GCC GTG GTG ATG GTG ATG ATG GCC GCT GCC
RT-Lib	G ACT ACA AGG ACG ACG ACG ACA AG**T AG**N NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNG GCA GCG GCC ATC ATC ACC ATC

RT-For	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG ACG ACG ACA AGT AG
RT-Rev	ATA GCC AGA TCC AGA CAT TCC CAT AGA ACC GCC GTG GTG ATG GTG ATG ATG GCC GCT GCC
AVA95	TTT TTT TTT TTT ATA GCC AGA TCC
AVA96	TTT TTT TTT TTT NAT AGC CAG ATC C
AVA111	ATA AAC ACA CAT AAA CAA ACA AAG AAT TCA TGT CTA AAG GTG AAG AAT TAT TCA CTG GTG
AVA112	GCT AGC TTT GTA CAA TTC ATC CAT ACC ATG GGT AAT ACC
AVA118	CAT ATT ATC TTC TTC ACC TTT TGA AAC CAT GAC GTC TCC AGA CAT TCC CAT AGA ACC GCC
AVA119	ATT ACC CAT GGT ATG GAT GAA TTG TAC AAA GCT AGC GGC AGC GGC GAC TAC
AVA136	AGC TAG CGG CAG CGG CGA CTA CAA GGA CGA CGA CGA CAA GTA G
AVA314	TTC CCT ACA CGA CGC TCT TCC GAT CTT ACA AGG ACG ACG ACG ACA AGT AG
AVA315	TTC CCT ACA CGA CGC TCT TCC GAT CTN TAC AAG GAC GAC GAC GAC AAG TAG
AVA316	TTC CCT ACA CGA CGC TCT TCC GAT CTN NTA CAA GGA CGA CGA CGA CAA GTA G
AVA319	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC
AVA320	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC ATG GTG ATG ATG GCC GCT GCC
AVA321	CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CAG ACG TGT GCT C

Table 2-8. Sequences of readthrough and control selection constructs.

Construct	Sequence
CT-Lib	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CA*A TGG ACT ACA AGG ACG ACG ACG ACA AG*T ACN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNG GCA GCG GCC ATC ATC ACC ATC ACC ACG GCG GTT CTA TGG GAA TGT CTG GAT CTG GCT AT
Inactive-RT	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CA*A TGG ACT ACA AGG ACG ACG ACG ACA AG*A CCC TAG ATG AC**T AG**C ACC GAG AAC CCG GCG CCA CGC AAT GGA ACG TCC TTA ACT CCG GCA GGC AAT TAA AGG GAA CGT ATG GCA GCG GCC ATC ATC ACC ATC ACC ACG GCG GTT CTA TGG GAA TGT CTG GAT CTG GCT AT
MLV	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CA*A TGG ACT ACA AGG ACG ACG ACG ACA AG*A CCC TAG ATG AC**T AG**G GAG GTC AGG GTC AGG AGC CCC CCC CTG AAC CCA GGA TAA CCC TCA AAG TCG GGG GGC AAC CCG TCG GCA GCG GCC ATC ATC ACC ATC ACC ACG GCG GTT CTA TGG GAA TGT CTG GAT CTG GCT AT
RT-Lib	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CA*A TGG ACT ACA AGG ACG ACG ACG ACA AG***T AG**N NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNN NNG GCA GCG GCC ATC ATC ACC ATC ACC ACG GCG GTT CTA TGG GAA TGT CTG GAT CTG GCT AT

Underlined – T7 promoter. Italicized – FLAG tag. Bold – Stop codon.

2.6 References

1. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).
2. Zaher, H. S. & Green, R. Fidelity at the Molecular Level: Lessons from Protein Synthesis. *Cell* **136,** 746–762 (2009).
3. Gesteland, R. F. & Atkins, J. F. Recoding: Dynamic Reprogramming of Translation. *Annu. Rev. Biochem.* **65,** 741–768 (1996).
4. Baranov, P. V., Atkins, J. F. & Yordanova, M. M. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nat. Rev. Genet.* **16,** 517–529 (2015).
5. *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. (Springer New York, 2010). at <http://link.springer.com/10.1007/978-0-387-89382-2>
6. Bertram, G., Innes, S., Minella, O., Richardson, J. & Stansfield, I. Endless possibilities: translation termination and stop codon recognition. *Microbiol. Read. Engl.* **147,** 255–269 (2001).
7. Firth, A. E. & Brierley, I. Non-canonical translation in RNA viruses. *J. Gen. Virol.* **93,** 1385–1409 (2012).
8. Loughran, G. *et al.* Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* **42,** 8928–8938 (2014).
9. Stiebler, A. C. *et al.* Ribosomal Readthrough at a Short UGA Stop Codon Context Triggers Dual Localization of Metabolic Enzymes in Fungi and Animals. *PLoS Genet* **10,** e1004685 (2014).
10. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife* **2,** e01179 (2013).
11. Freitag, J., Ast, J. & Bölker, M. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature* **485,** 522–525 (2012).
12. Wills, N. M., Gesteland, R. F. & Atkins, J. F. Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci.* **88,** 6991–6995 (1991).

13. Alam, S. L., Wills, N. M., Ingram, J. A., Atkins, J. F. & Gesteland, R. F. Structural studies of the RNA pseudoknot required for readthrough of the gag-termination codon of murine leukemia virus1. *J. Mol. Biol.* **288,** 837–852 (1999).
14. Napthine, S., Yek, C., Powell, M. L., Brown, T. D. K. & Brierley, I. Characterization of the stop codon readthrough signal of Colorado tick fever virus segment 9 RNA. *RNA* **18,** 241–252 (2012).
15. Green, L., Houck-Loomis, B., Yueh, A. & Goff, S. P. Large ribosomal protein 4 increases efficiency of viral recoding sequences. *J. Virol.* **86,** 8949–8958 (2012).
16. Williams, I., Richardson, J., Starkey, A. & Stansfield, I. Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.* **32,** 6605–6616 (2004).
17. Jungreis, I. *et al.* Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res.* **21,** 2096–2113 (2011).
18. Namy, O., Hatin, I. & Rousset, J.-P. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.* **2,** 787–793 (2001).
19. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).
20. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249,** 505–510 (1990).
21. Robertson, D. L. & Joyce, G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344,** 467–468 (1990).
22. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci.* **94,** 12297–12302 (1997).
23. Wilson, D. S. & Szostak, J. W. In Vitro Selection of Functional Nucleic Acids. *Annu. Rev. Biochem.* **68,** 611–647 (1999).
24. Jenison, R. D., Gill, S. C., Pardi, A. & Polisky, B. High-resolution molecular discrimination by RNA. *Science* **263,** 1425–1429 (1994).
25. Berens, C., Thain, A. & Schroeder, R. A tetracycline-binding RNA aptamer. *Bioorg. Med. Chem.* **9,** 2549–2556 (2001).

26. Weigand, J. E. *et al.* Screening for engineered neomycin riboswitches that control translation initiation. *RNA* **14,** 89–97 (2008).

27. Joyce, G. F. Directed Evolution of Nucleic Acid Enzymes. *Annu. Rev. Biochem.* **73,** 791–836 (2004).

28. Frankel, A. & Roberts, R. W. In vitro selection for sense codon suppression. *RNA N. Y. N* **9,** 780–786 (2003).

29. Liu, C. C. & Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **79,** 413–444 (2010).

30. Liu, R., Barrick, J. E., Szostak, J. W. & Roberts, R. W. Optimized synthesis of RNA-protein fusions for in vitro protein selection. *Methods Enzymol.* **318,** 268–293 (2000).

31. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. & Szostak, J. W. Constructing high complexity synthetic libraries of long ORFs using In Vitro selection. *J. Mol. Biol.* **297,** 309–319 (2000).

32. Huang, J., Brieba, L. G. & Sousa, R. Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. *Biochemistry (Mosc.)* **39,** 11571–11580 (2000).

33. Houck-Loomis, B. *et al.* An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature* **480,** 561–564 (2011).

34. Seelig, B. mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. *Nat. Protoc.* **6,** 540–552 (2011).

35. Li, G. & Rice, C. M. The signal for translational readthrough of a UGA codon in Sindbis virus RNA involves a single cytidine residue immediately downstream of the termination codon. *J. Virol.* **67,** 5062–5067 (1993).

36. Skuzeski, J. M., Nichols, L. M., Gesteland, R. F. & Atkins, J. F. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *J. Mol. Biol.* **218,** 365–373 (1991).

37. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the Genetic Code of Escherichia coli. *Science* **292,** 498–500 (2001).

38. Johnson, D. B. F. *et al.* RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat. Chem. Biol.* **7,** 779–786 (2011).

39. *Current Protocols in Molecular Biology*. at <http://onlinelibrary.wiley.com/book/10.1002/0471142727>

40. Keppler-Ross, S., Noffz, C. & Dean, N. A New Purple Fluorescent Color Marker for Genetic Studies in Saccharomyces cerevisiae and Candida albicans. *Genetics* **179,** 705–710 (2008).

41. Pirakitikulr, N., Ostrov, N., Peralta-Yahya, P. & Cornish, V. W. PCRless library mutagenesis via oligonucleotide recombination in yeast. *Protein Sci. Publ. Protein Soc.* **19,** 2336–2346 (2010).

Chapter 3

Reprogramming Eukaryotic Translation with Ligand-Responsive Synthetic RNA Switches

---

*The contents of this chapter will be published in:

A.V. Anzalone, V.W. Cornish, et al "Reprogramming Eukaryotic Translation with Ligand-Responsive Synthetic RNA Switches." *In preparation*.

3.0 Chapter Outlook

Protein translation in eukaryotes is reprogrammed through diverse molecular mechanisms to expand the coding capacity of individual genes. One such mechanism termed -1 programmed ribosomal frameshifting (-1 PRF) utilizes cis-acting signals within the mRNA to redirect ribosomes into a new translation reading frame that gives rise to an alternative protein product. In this chapter, we establish -1 PRF as a powerful framework for engineering small molecule responsive RNA switches that control protein synthesis in eukaryotes. First, we implemented our *in vitro* selection for translation reprogramming developed in Chapter 2 to identify efficient -1 PRF stimulatory elements and characterized the selection products by high-throughput sequencing. Then, we demonstrate the construction of ligand responsive -1 PRF switches by coupling -1 PRF stimulatory elements to RNA aptamers using rational design and *in vivo* directed evolution. Our -1 PRF switches can control the relative stoichiometry of two distinct protein outputs from a single mRNA and demonstrate consistent ligand response across whole populations of cells. We further applied -1 PRF switches to build condensed logic gates and an apoptosis module in yeast. Together, these results showcase the potential for engineering RNA to reprogram translation, opening new opportunities for scalable and precise control of protein synthesis in eukaryotes

3.1 Introduction

The ribosome coordinates the biosynthesis of proteins from mRNA templates according to a standard translational program. While this process typically proceeds with high fidelity[1], in some cases, the translational program is momentarily altered in order to change the protein output of a gene[2,3]. This 'reprogramming' endows the translation apparatus with expanded synthetic capabilities, enabling the expression of proteins containing non-canonical amino acids (such as selenocysteine or pyrrolysine) or the regulated expression of multiple distinct protein products from a single mRNA transcript[4]. Some forms of translation reprogramming have been directly adopted for biotechnology, including internal ribosome entry sites (IRES)[5] and co-translational cleaving 2A peptides[6]. Moreover, substantial effort has been directed toward engineering translation components for the redefinition of stop codons to specify unnatural amino acids in living cells[7]. While significant progress has been made in these areas, other modes of translation reprogramming remain largely unexplored despite their potential applications for synthetic biology.

Several reprogramming mechanisms utilize cis-acting RNA elements to signal the recoding event at a specific location along the mRNA, offering attractive opportunities to engineer RNA devices that regulate protein synthesis. Encouragingly, other RNA-based gene-regulatory frameworks have recently emerged as powerful tools for engineering biological systems[8]. Over two decades of SELEX and related *in vitro* selection technologies[9–11] have yielded synthetic RNA molecules that bind to diverse ligands[12,13] and catalyze chemical reactions. Engineered RNA devices, particularly allosteric ribozymes[14–18], have been applied to synthetic biology for cellular computation[19], gene

regulation[20], and controlling cell phenotype[21,22]. Much of this success may be attributed to the modularity of RNA device construction, which could be broadly applicable beyond previously explored classes of RNA devices.

We identified -1 Programmed Ribosomal Frameshifting (-1 PRF) as a potentially modular RNA mechanism for developing new gene-regulatory tools. Eukaryotic -1 PRF signals are composed of two general features: (i) a heptanucleotide slippery site where the frameshift event occurs, with the general sequence X-XXY-YYZ (dashes demarcate codons in the original frame; X can be any nucleotide; Y can be A or U; and Z can be A, C or U); and (ii) a downstream stimulatory RNA structure, typically a hairpin or pseudoknot[23]. When encountering a -1 PRF signal in an mRNA, a fraction of translating ribosomes slip back by a single nucleotide, placing the translation apparatus in the -1 reading frame. This, in turn, alters the amino acid composition of the polypeptide that is synthesized downstream of the frameshift site.

-1 PRF is especially common in retroviruses, including HIV, where it serves to establish a precise stoichiometry of Gag and Gag-Pol proteins[24]. Regardless of variation in mRNA transcript levels or translational activity, the stoichiometry of frameshift to non-frameshift protein products remains constant for a given -1 PRF signal. Thus, -1 PRF is an attractive mechanism to set the relative abundance of multiple proteins by co-encoding them on a single transcript. While viral -1 PRF signals have fixed activities, we were inspired by the natural polyamine sensing antizyme +1 frameshift signal[25] to engineer ligand responsive -1 PRF devices (**Fig. 3-1a**). It was previously demonstrated *in vitro* and in living cells that certain prokaryotic transcriptional riboswitches can serve as frameshift stimulatory elements and trigger -1 PRF in response to their cognate

ligands[26,27]. However, a general strategy for constructing synthetic -1 PRF devices that can be controlled by orthogonal ligands has not yet been developed.

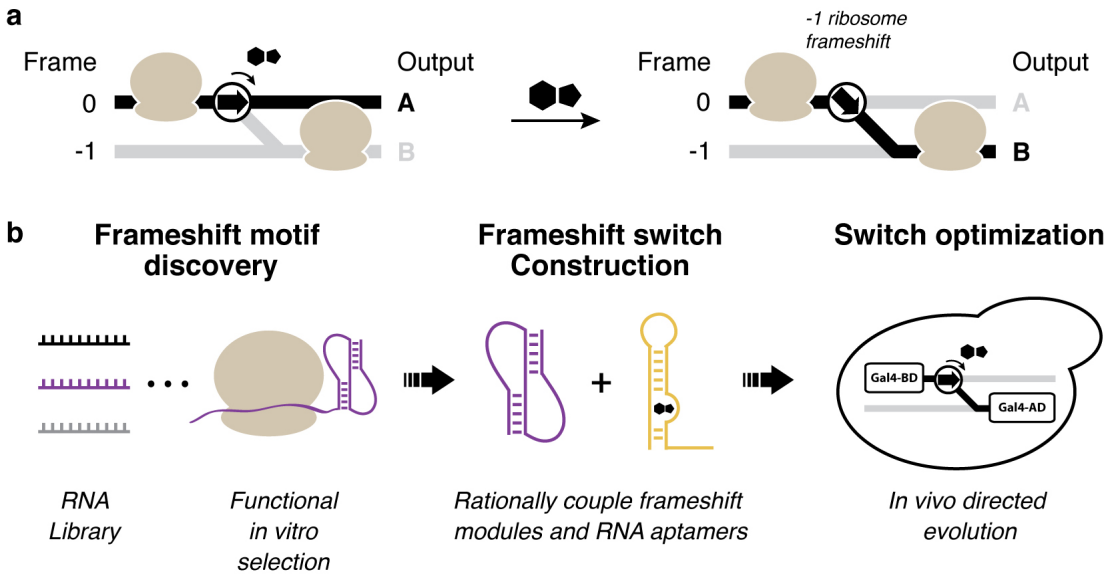


Figure 3-1. Building ligand responsive -1 PRF switches. (**a**) Gene control concept. The protein output of an mRNA is dictated by the translation reading frame. -1 PRF switch devices direct the ribosome's translation reading frame depending on the presence or absence of a ligand. (**b**) Methodological approach towards building -1 PRF switches. Active frameshift stimulatory elements are discovered from large RNA libraries using a functional *in vitro* selection. Frameshift stimulatory elements are then coupled to RNA aptamer modules by rational design to create ligand responsive -1 PRF switches. Lastly, devices can be optimized by *in vivo* directed evolution in yeast using a -1 PRF regulated transcriptional activator as a selectable output.

Here, we establish a platform for engineering ligand-responsive -1 PRF switch devices from synthetic components and apply them for *in vivo* regulation of protein synthesis. We describe a functional *in vitro* selection for -1 PRF stimulatory element discovery, combining the combinatorial complexity of a classical *in vitro* selection with the biochemical complexity of a cell lysate. We further leverage rational design and *in vivo* directed evolution to construct robust ligand-responsive -1 PRF switches (**Fig. 3-1b**).

Our devices display robust behavior by balancing protein output stoichiometry, enabling the assembly of logic gates and a phenotypic control switch that responds with precision at the individual cell level.

3.2 Results

3.2.1 An *in vitro* selection for -1 programmed ribosomal frameshifting

To establish a -1 PRF toolkit for synthetic biology, we set out to compile a collection of robust RNA frameshift stimulatory elements that are suitable for downstream applications. Ideally, these elements are short in length, have well-defined structures, trigger high efficiency -1 PRF, and are readily exchanged with other frameshift elements of similar size and composition. However, natural -1 PRF stimulatory elements are diverse in both size and sequence, and in some cases, require >3000 nucleotides for frameshift activity. Moreover, natural -1 PRF signals stimulate frameshifting at levels that are tuned for optimal viral replication (typically 5-10%), not maximal efficiency[3]. The above limitations place constraints on the scope of available parts, the modularity for engineering, and the achievable dynamic range of ligand-responsive devices. Therefore, we developed a selection strategy to discover novel -1 PRF stimulatory elements from large libraries of sequence variants derived from a uniformly compact scaffold.

The directed evolution approach that we designed exploits mRNA display, a well-established *in vitro* selection technology[28]. During *in vitro* translation of mRNA-display templates, the amplifiable genotype of the mRNA is connected to the phenotype of the polypeptide through a critical puromycin ligation reaction. While mRNA-display is conventionally employed to select for functional proteins and peptides, we repurposed it

to select for the translation reprogramming activity of an mRNA (**Fig. 3-2**). Due to the strong distance dependence of the puromycin reaction[29], mRNA-peptide fusion occurs only if the ribosome translates the entire mRNA transcript. As a result, upstream stop codons that terminate translation prevent mRNA-peptide fusion from occurring[30]. We exploited this specificity of puromycin reactivity to differentiate frameshift stimulating sequences from inactive sequences.

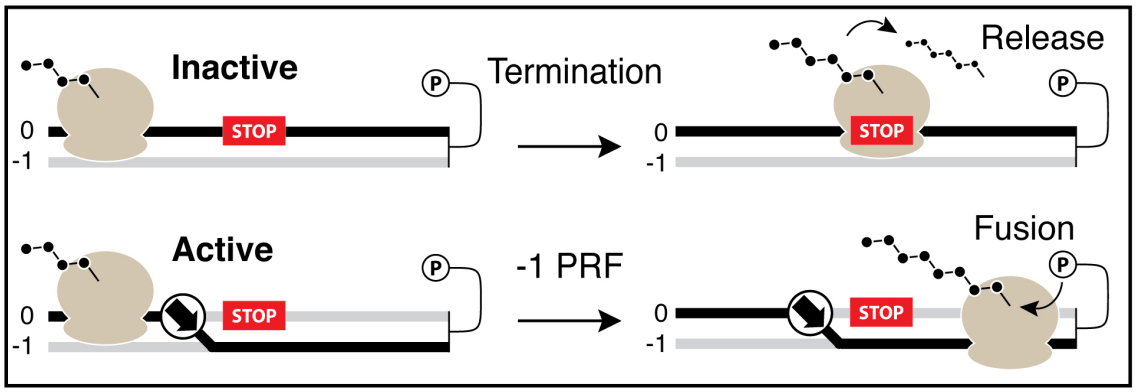


Figure 3-2. Translation reprogramming selection principle. mRNA-display templates form mRNA-peptide fusions if the ribosome translates the entire transcript. Ribosomes that terminate translation upstream of the designated fusion point will fail to produce mRNA-peptide fusions (upper). Frameshifting enables bypass of encoded stop codons to produce mRNA-peptide fusions (lower).

An mRNA library was designed such that only active -1 PRF signals promote mRNA-peptide fusion formation. Starting from a prokaryotic riboswitch scaffold[26], 14 nucleotides were randomized to generate 2.7×10^8 sequence variants (see **Appendix A.1**). Notably, a library of this size is easily accommodated by the mRNA-display technology, which allows for upwards of 10^{14} input sequences[29]. To enrich active -1 PRF stimulatory elements, the library was encoded downstream of the heptanucleotide slippery site U-UUA-AAC and an in-frame UAG termination codon (**Fig. 4-3**). After three cycles of *in*

vitro selection, assaying of selection products in a dual-fluorescent protein (dual-FP) reporter in *S. cerevisiae* revealed enrichment for active frameshift stimulating elements (**Fig. 3-4** and **Fig. 3-5**).

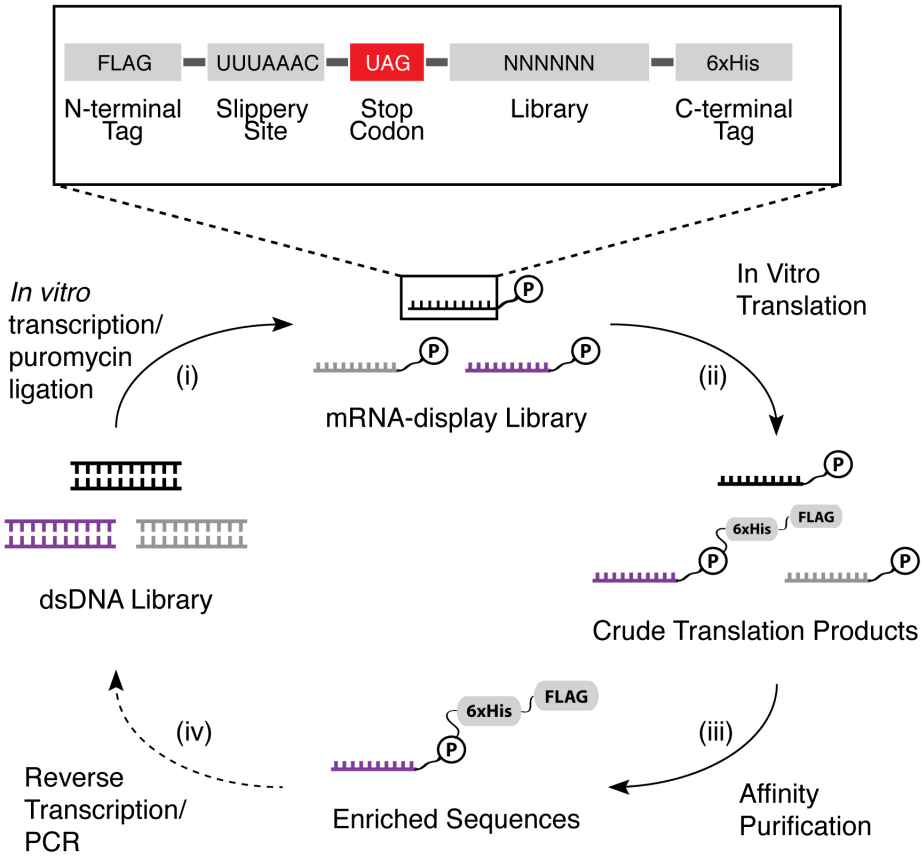


Figure 3-3. *In vitro* selection for -1 PRF stimulatory elements. (**a**) The mRNA display selection. The display construct encodes an N-terminal FLAG tag, a heptanucleotide slippery site, an in-frame stop codon, the stimulatory element library, and a C-terminal hexahistidine tag encoded in the -1 frame. The selection cycle comprises four stages: (i) DNA is *in vitro* transcribed and ligated to the puromycin adaptor; (ii) mRNA display templates are translated *in vitro* in rabbit reticulocyte lysate; (iii) mRNA-peptide fusions are isolated from non-fused RNA by affinity purification of the peptide tags; (iv) enriched sequences are reverse transcribed and PCR amplified for subsequent rounds of selection.

Flow cytometry of the pooled selection products within dual-FP reporter plasmids showed significant enrichment of active frameshift stimulators, displaying *in vivo* efficiencies of up to 30% (**Fig. 3-5**). Moreover, flow cytometry of individual clones revealed that the ratio of fluorescent proteins remained constant for a given population of cells harboring the same -1 PRF signal, irrespective of total protein synthesis (**Fig. 3-4**). As a result, two populations of yeast with frameshifting efficiencies that differ by only 3- to 4-fold are highly resolvable, despite overall expression levels that span several orders of magnitude.

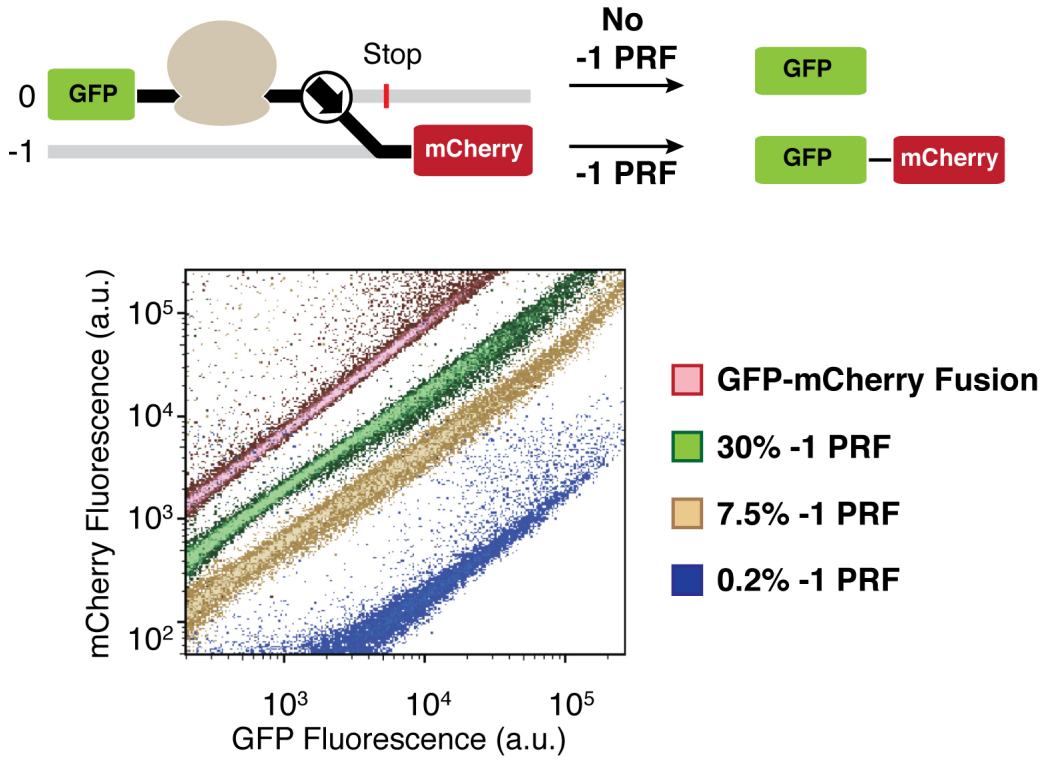


Figure 3-4. Dual-FP reporter assay of -1 PRF efficiency in *S. cerevisiae*. The frameshift variant is cloned between green fluorescent protein (GFP) and the red fluorescent protein mCherry. The ratio of fluorescence signals reflects bulk -1 PRF efficiency. Flow cytometry of individual clones demonstrates the stable stoichiometry produced by -1 PRF signals across various expression levels.

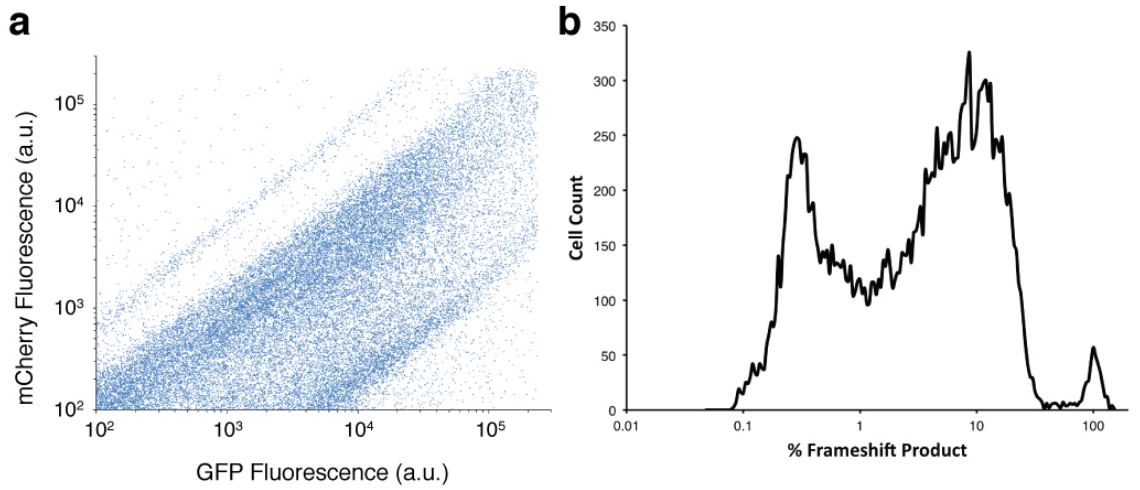


Figure 3-5. Characterization of *in vitro* selection products by flow cytometry. (**a**) Dot plot and (**b**) histogram of bulk *in vitro* selection products cloned into the dual-FP reporter.

The *in vitro* selection products were submitted for next-generation sequencing (NGS) to characterize the landscape of frameshift stimulatory elements. A computational pipeline was implemented to identify promising sequence motifs for downstream engineering applications. Sequences were grouped on the basis of compatibility with different geometries of H-type pseudoknots, and also clustered on their primary sequence identity (see **Appendix A.1**). PK enrichment was assessed by explicitly computing the initial probability distribution over the space of PK structures, and comparing with the empirical distribution observed post-selection (**Fig. 3-6**). Following motif grouping, comparative analysis and secondary structure prediction algorithms (pKiss[31]) were used to further support secondary structure assignment. Additionally, the differential abundance of variants within a motif was used to identify nucleotide preferences at variable sites and mutation intolerant positions (**Appendix A.1**).

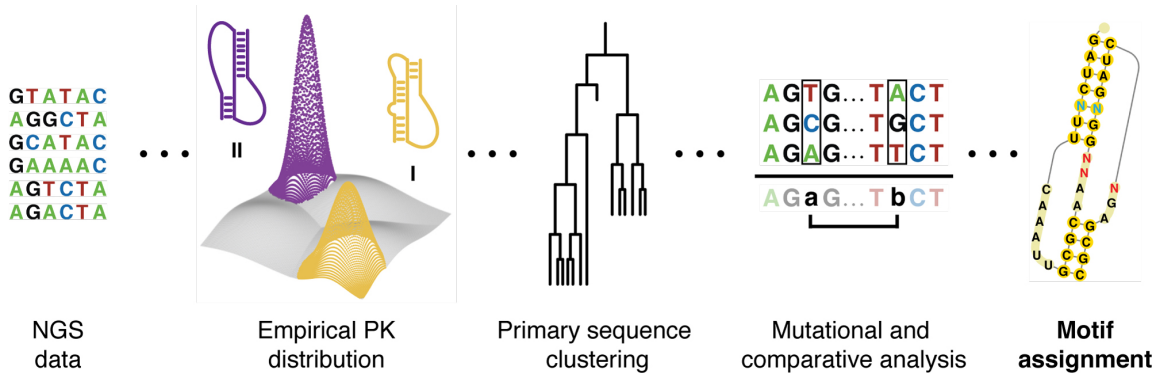


Figure 3-6. NGS analysis of in vitro selected -1 PRF stimulators. Sequences were analyzed for their compatibility with various H-type pseudoknot base-pairing geometries, then clustered based on primary sequence identity. Comparative analysis was applied to support base pair predictions, and mutational analysis identified sequences constraints on the motif. See **Appendix A.1** for further details.

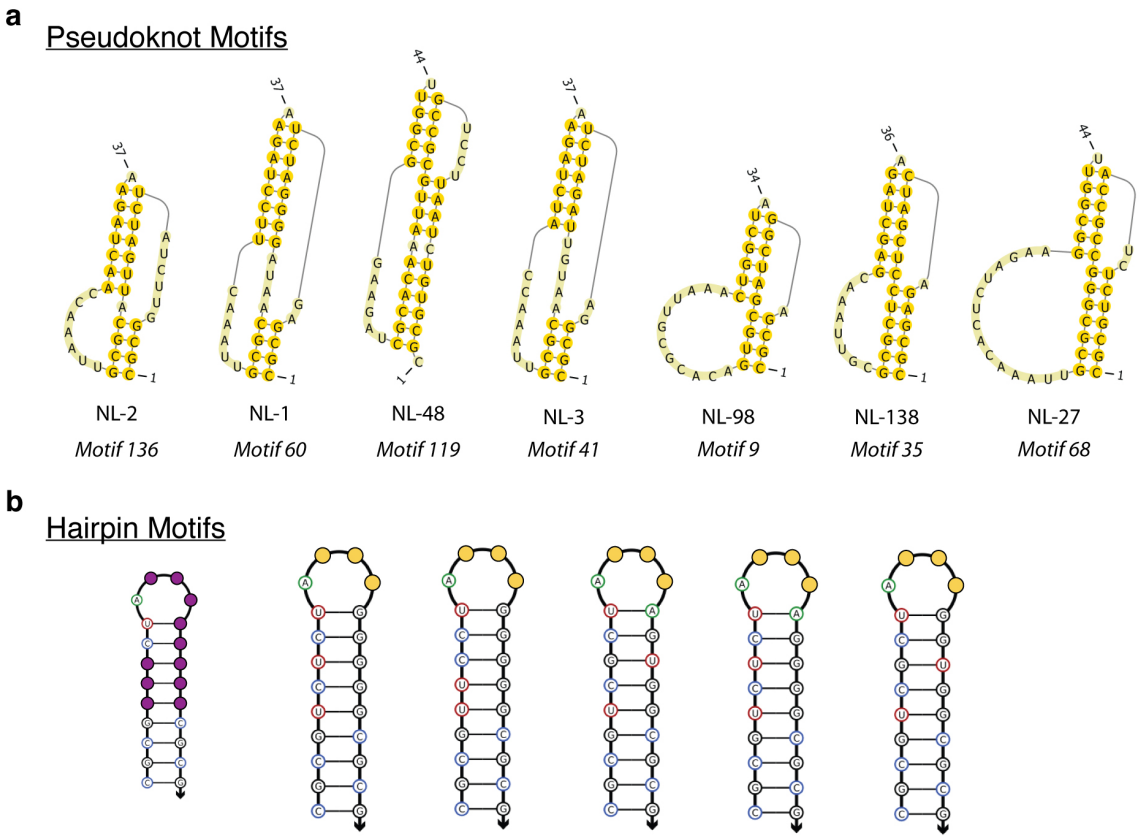


Figure 3-7. Frameshift stimulatory motifs. (**a**) H-type pseudoknot and (**b**) hairpin motifs identified by NGS (purple circles indicate nucleotide positions that were randomized in the starting library).

3.2.2 Rational design and *in vivo* optimization of ligand responsive -1 PRF switches

To engineer ligand responsive devices, we rationally coupled our -1 PRF stimulatory elements to small molecule binding RNA aptamers. For our stimulatory element, we chose sequence NL-2 from the *in vitro* selection because it displayed high -1 PRF efficiency in yeast (30%), was the second most abundant sequence in the NGS, and has a confidently predicted pseudoknot fold (**Fig. 3-7**). As aptamer domains, we chose the theophylline[12] and neomycin[32] binding aptamers based on their previous success for *in vivo* applications.

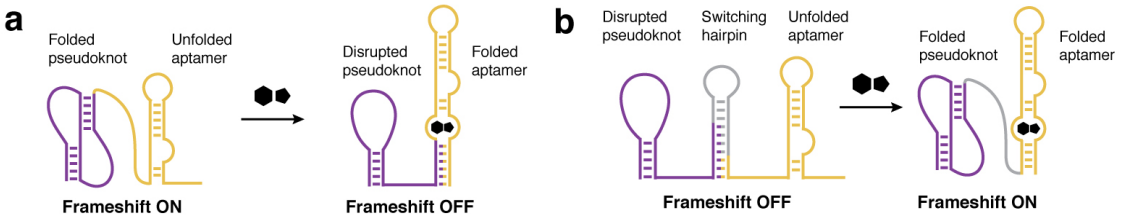


Figure 3-8. Rational design of frameshift switches. (**a**) OFF-switch design. In the absence of ligand, the stimulatory pseudoknot (purple) is energetically dominant, producing high frameshift levels. Ligand binding induces aptamer (gold) folding, which disrupts the pseudoknot structure leading to lowered frameshift levels. (**b**) ON-switch design. A switching hairpin (gray) is installed to disrupt the pseudoknot and lower basal frameshifting. In the presence of ligand, the aptamer folds and destabilizes the switching hairpin, allowing the pseudoknot to re-fold and restore frameshift activity.

In OFF-switches (**Fig. 3-8a**), the RNA aptamer and NL-2 pseudoknot sequences overlap creating competition for folding. In the absence of ligand, the active pseudoknot predominates and stimulates high -1 PRF activity. However, in the presence of ligand, the folded aptamer is stabilized by ligand binding energy and disrupts the NL-2 pseudoknot, resulting in lowered -1 PRF efficiency. We designed several constructs by varying the

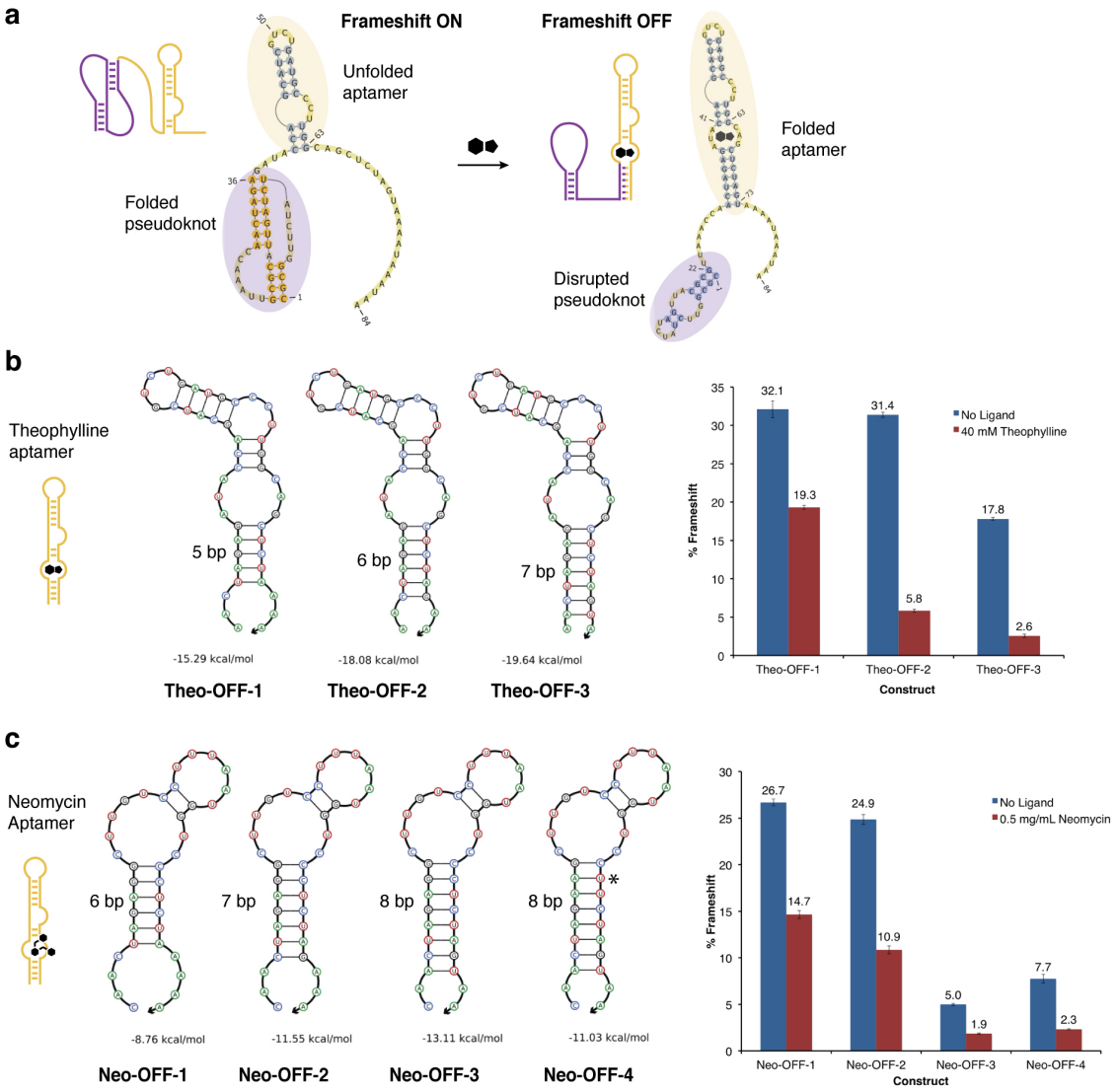


Figure 3-9. Design and characterization of OFF switch devices. (**a**) General scheme depicting the intended structural rearrangements. A representative theophylline responsive OFF switch sequence is shown. (**b**) Optimization of a theophylline responsive OFF-switch. Increasing the length of the theophylline aptamer stem leads to increasing disruption of the stimulatory pseudoknot. Frameshift levels, both in the absence and presence of ligand, are inversely correlated with aptamer thermodynamic stability. (**c**) Efforts to optimize a neomycin responsive OFF-switch. As with the theophylline responsive switches, in general, frameshift levels are inversely correlated to the thermodynamic stability of the aptamer. However, this relationship is not as direct, and this method was unable to provide a switch that possessed ideal -1 PRF levels and ligand responsiveness. Instead, the neomycin OFF-switch was optimized by in vivo directed evolution (see **Fig. 3-13**).

length and composition of the aptamer stems. As a general trend, we found that increasing the thermodynamic stability of the aptamer lowered frameshift activity (**Fig. 3-9**). This simple approach led to the discovery of a high-performing theophylline OFF switch that displays a 7-fold reduction in -1 PRF in the presence of ligand.

Unfortunately, rational optimization of aptamer stability did not result in adequate neomycin OFF-switch activity. Therefore, we devised an *in vivo* directed evolution platform for optimizing switch performance that exploits the unique output of -1 PRF (**Fig. 3-10**). Briefly, -1 PRF devices were inserted between the open reading frames of a Gal4 minimal DNA binding domain (BD) and a Gal4 minimal activation domain (AD)[33]. Overall, expression of Gal4 responsive genes in a yeast two-hybrid strain[34] is a function of the Gal4(BD) / Gal4(BD-AD) ratio, which in turn is dictated by -1 PRF efficiency (**Fig. 3-10**). Additionally, we also demonstrated that functional expression of selectable markers in the yeast two-hybrid strain is dependent of frameshift efficiency. In growth assays (**Fig. 3-11**), fixed -1 PRF stimulators of varying efficiency were cloned between the Gal4(BD) and Gal4(AD), then evaluated for growth in the positive selection conditions and counterselection conditions. We found that higher efficiency -1 PRF stimulators grew better in positive selection conditions, but were very sensitive to 5-FOA under counterselection conditions.

We designed a library of neomycin OFF-switches by varying nucleotides in the aptamer's terminal stem region. After one round of positive selection in the absence of ligand and counter-selection in the presence of neomycin, we identified multiple switches with improved performance (**Fig. 3-12**), the best of which demonstrated 5.5-fold change in -1 PRF in response to neomycin. Aside from directed evolution applications, this Gal4 system could be used for small molecule activated transcriptional regulation.

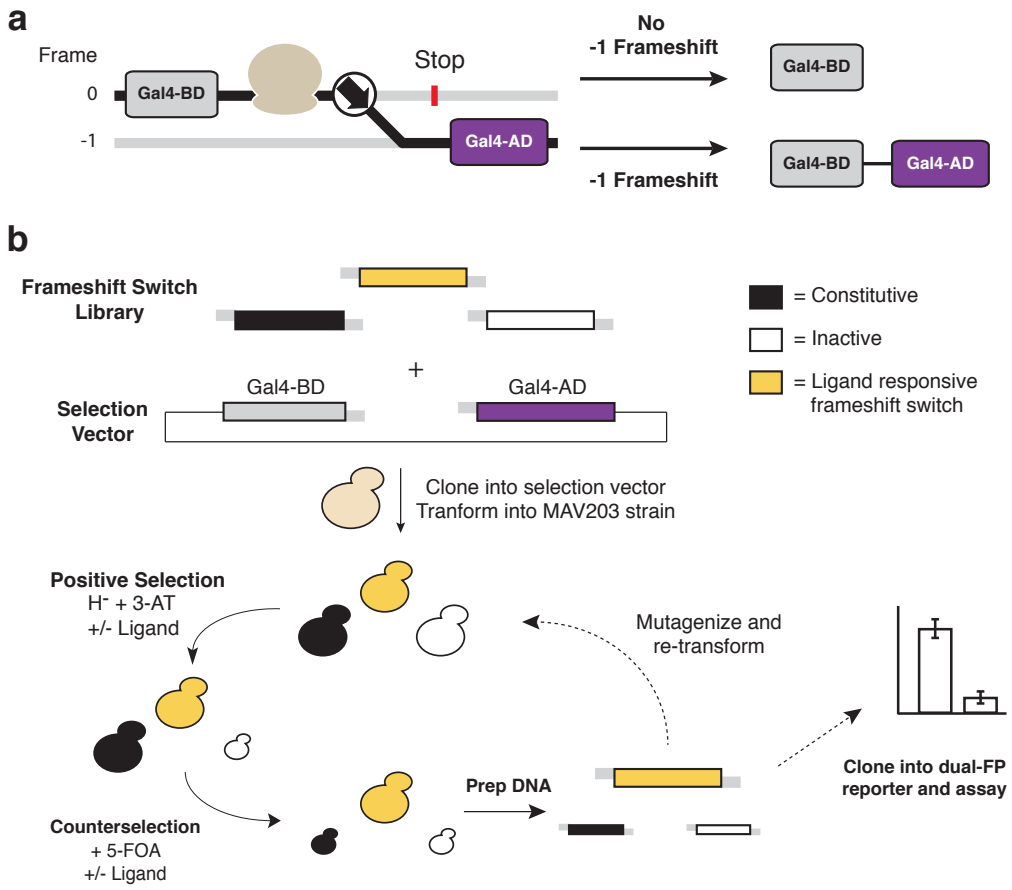


Figure 3-10. Directed evolution of -1 PRF switches in *S. cerevisiae*. (**a**) -1 PRF switches are encoded between a Gal4 minimal DNA binding domain (Gal4-BD) and a Gal4 minimal activation domain (Gal4-AD). Failure to frameshift at the switch results in translation termination to product a free Gal4-BD. Frameshifting allows for translation of the full minimal Gal4 transcription factor. Activation of Gal4 responsive genes is a function of Gal4-BD/Gal4-TF ratio and -1 PRF efficiency (see **Fig. 3-11** and **3-11**). (**b**). Selection scheme. A library of -1 PRF devices is cloned into the Gal4 selection vector by gap repair directly into the yeast 2-hybrid selection strain MAV203. The transformants are subjected to positive selection in media lacking histidine supplemented with the his3 inhibitor 3-aminotriazole (3-AT). For the selection of ON switches, media is also supplemented with the small molecule inducer. Counterselection is performed in minimal media containing 5-fluoroorotic acid (5-FOA) and the small molecule ligand in the case of OFF switch selections. After rounds of selection, DNA can be extracted from the culture and cloned into the dual-FP reporter for characterization, or mutagenized and retransformed for further selections.

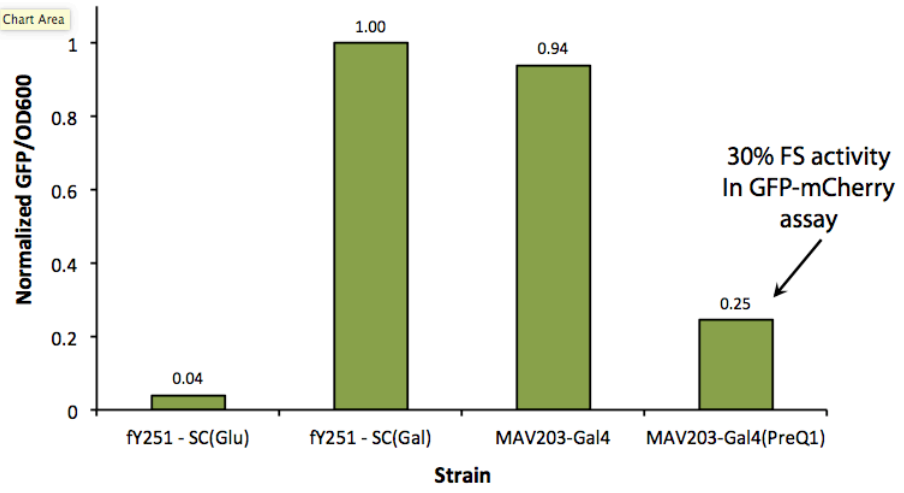


Figure 3-11. Assaying Gal4 activity with a GFP reporter.

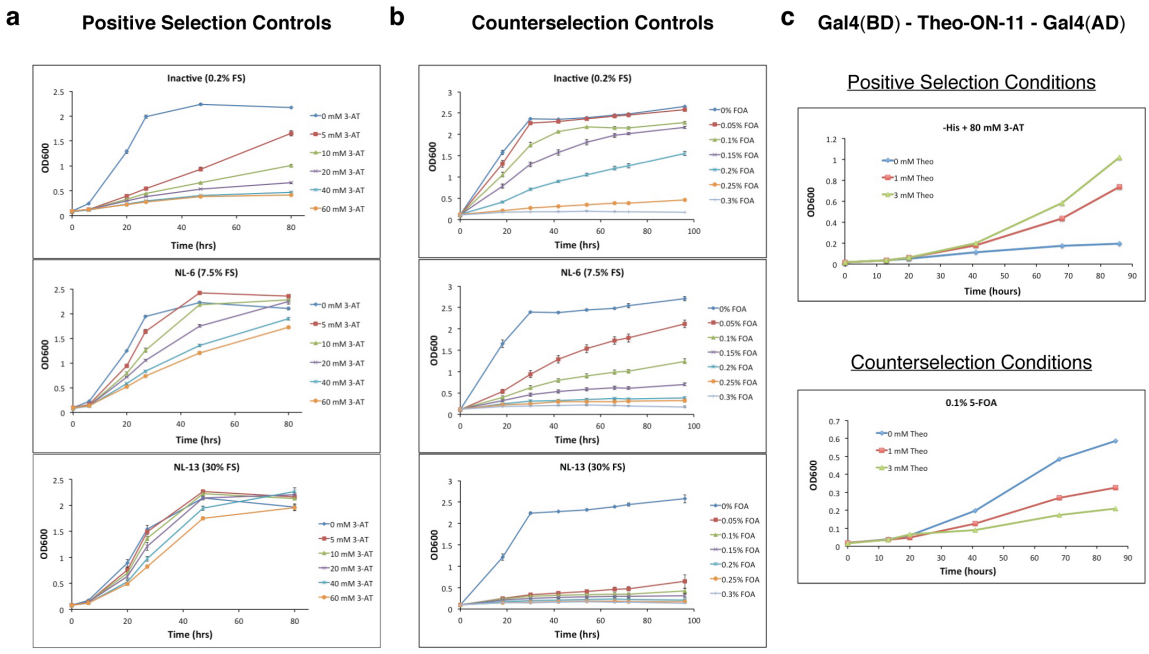


Figure 3-12. Growth assays of -1 PRF devices in the Gal4 selection construct. -1 PRF stimulators of varying efficiency were cloned into the Gal4 selection plasmid and assayed for growth in (**a**) positive selection conditions in His- minimal media containing varying amounts of 3-AT or (**b**) counterselection conditions media containing varying concentrations of 5-FOA. Higher efficiency -1 PRF stimulators grow better in positive selection, while lower efficiency -1 PRF stimulators grow better in the counterselection conditions. (**c**) A theophylline responsive ON-switch in the Gal4 selection. The construct shows theophylline-assisted growth in the positive selection conditions and growth inhibition by theophylline in the counterselection conditions.

For our ON-switches, a "switching hairpin" was introduced to compete with the NL-2 pseudoknot to reduce basal -1 PRF levels. By design, structural rearrangements stimulated by ligand binding serve to destabilize the switching hairpin, leading to coincident refolding of the pseudoknot and restoration of -1 PRF activity (**Fig. 3-8b**). Several constructs were designed with varying lengths and compositions of the aptamer stems and switching hairpins to tune the relative stabilities of the ON and OFF states. These constructs were assessed using RNA secondary structure prediction (NUPACK[35]) and tested experimentally in the dual-FP assay to correlate hairpin and aptamer stability to switch activity (**Fig. 3-14** and **Fig. 3-15**). With minimal optimization, this approach led to efficient ON-switches that respond to theophylline (5.9-fold) or neomycin (4.2-fold) (**Fig. 3-16a**).

When evaluating the theophylline responsive ON-switch Theo-ON-11 by flow cytometry, populations of cells exposed to theophylline can be distinguished from untreated cells with >99% accuracy (**Fig. 3-16b**). This highlights that the stoichiometric precision of -1 PRF is retained even in the context of a ligand responsive switch. Overall, these results demonstrate that rational design of RNA structural rearrangements controlled by small molecule ligands can be used to modulate -1 PRF and gene expression, and that devices can be quickly optimized by a Gal4-based *in vivo* directed evolution platform in yeast.

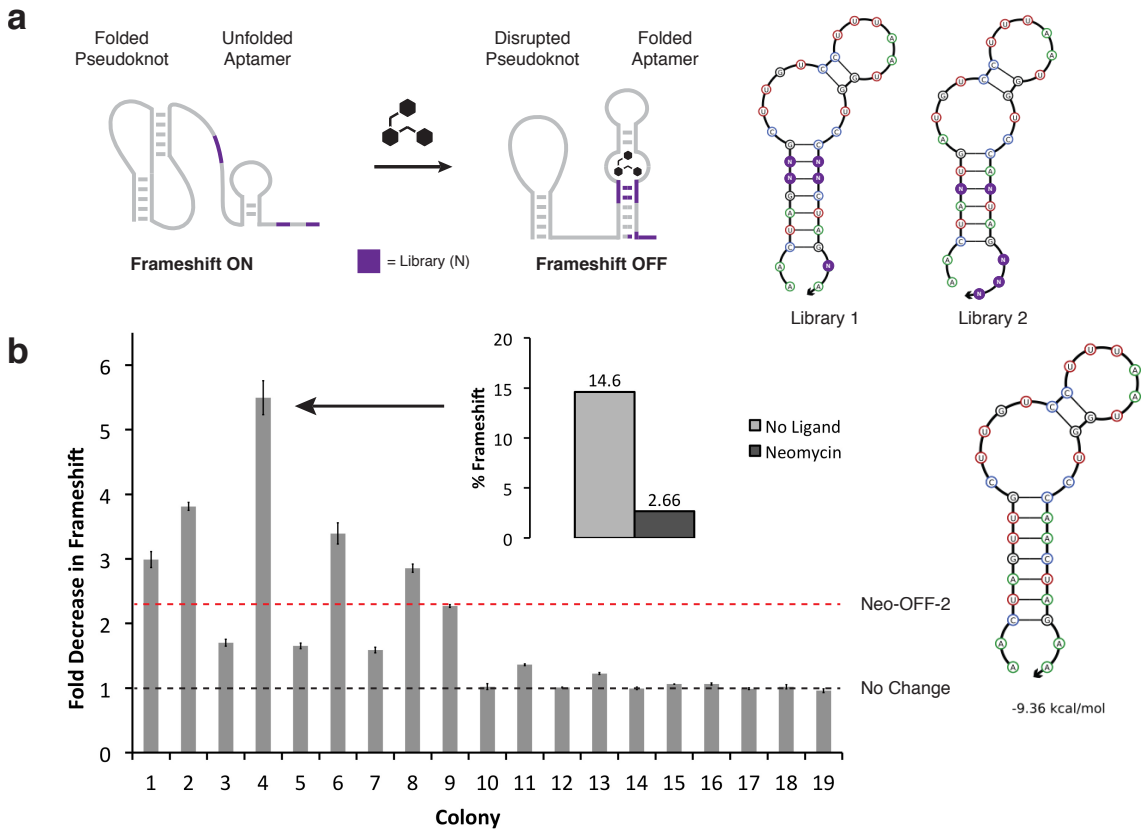


Figure 3-13. Optimization of a neomycin responsive OFF-switch by *in vivo* directed evolution. (**a**) Scheme of the neomycin OFF switch library targeting the aptamer stem. Two libraries, each containing 5 randomized nucleotides in the aptamer, were, cloned into the Gal4 selection vector and transformed into the yeast two-hybrid selection strain. (**b**) After one round of counterselection in the presence of neomycin and one round of positive selection in the absence of neomycin, selection products were cloned into the dual-FP reporter. Randomly sampled colonies were assayed for ligand responsiveness. One variant (aptamer structure depicted on the right) demonstrated significantly improved switch behavior.

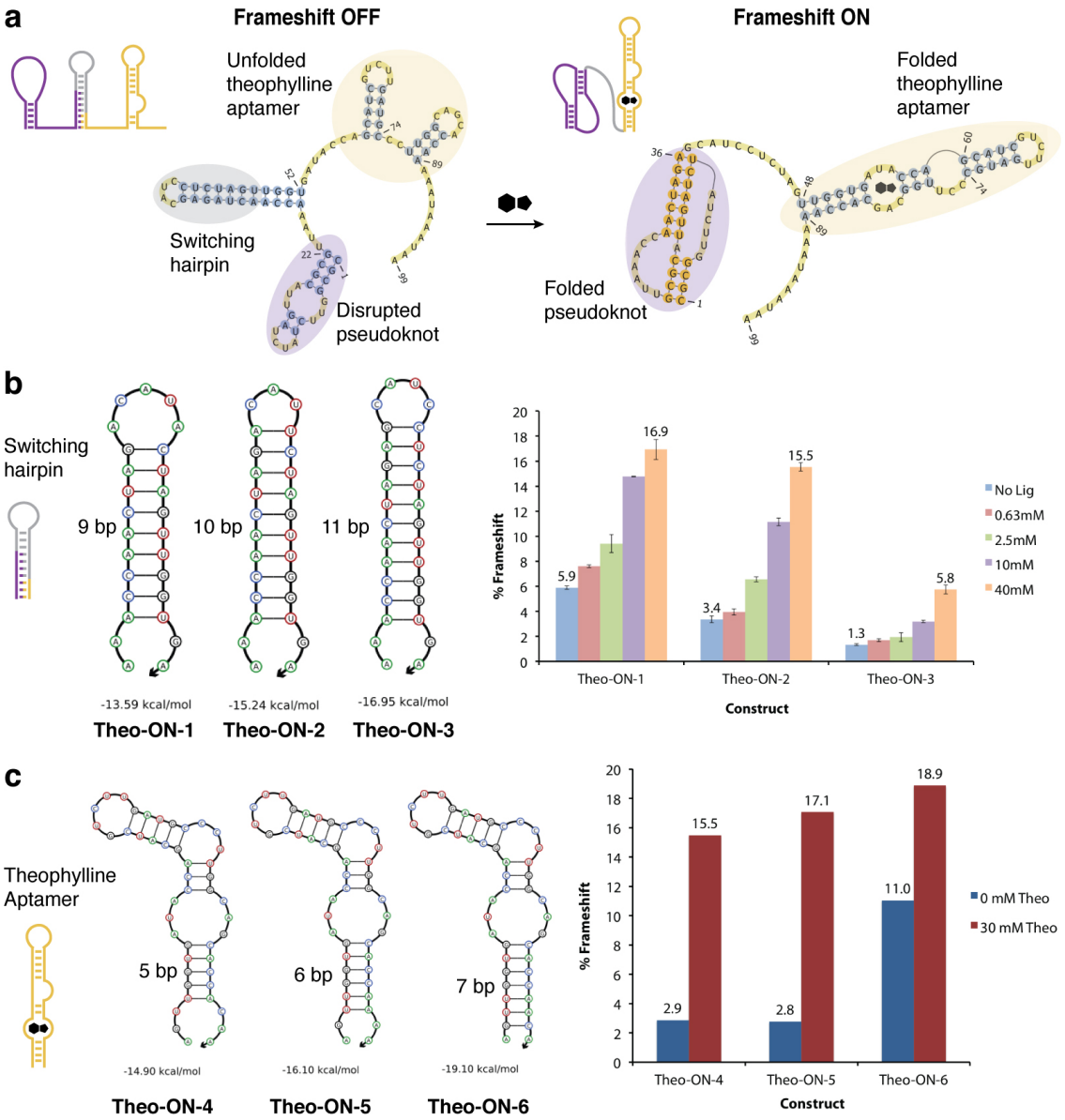


Figure 3-14. Design and characterization of theophylline responsive ON switches. (**a**) General scheme depicting the intended structural rearrangements induced by ligand binding. A representative theophylline responsive ON switch sequence is shown. (**b**) First, basal frameshifting levels were tuned down by inserting a switching hairpin to compete with pseudoknot folding. Basal frameshifting and ligand induced frameshifting is inversely correlated with switching hairpin length and thermodynamic stability. (**c**) Activation was optimized by adjusting the length of the aptamer stem. Frameshifting activity, in general, correlated with the stem length and thermodynamic stability of the aptamer. Incorporation of other thermodynamic parameters partially accounts for the discrepancy in this trend.

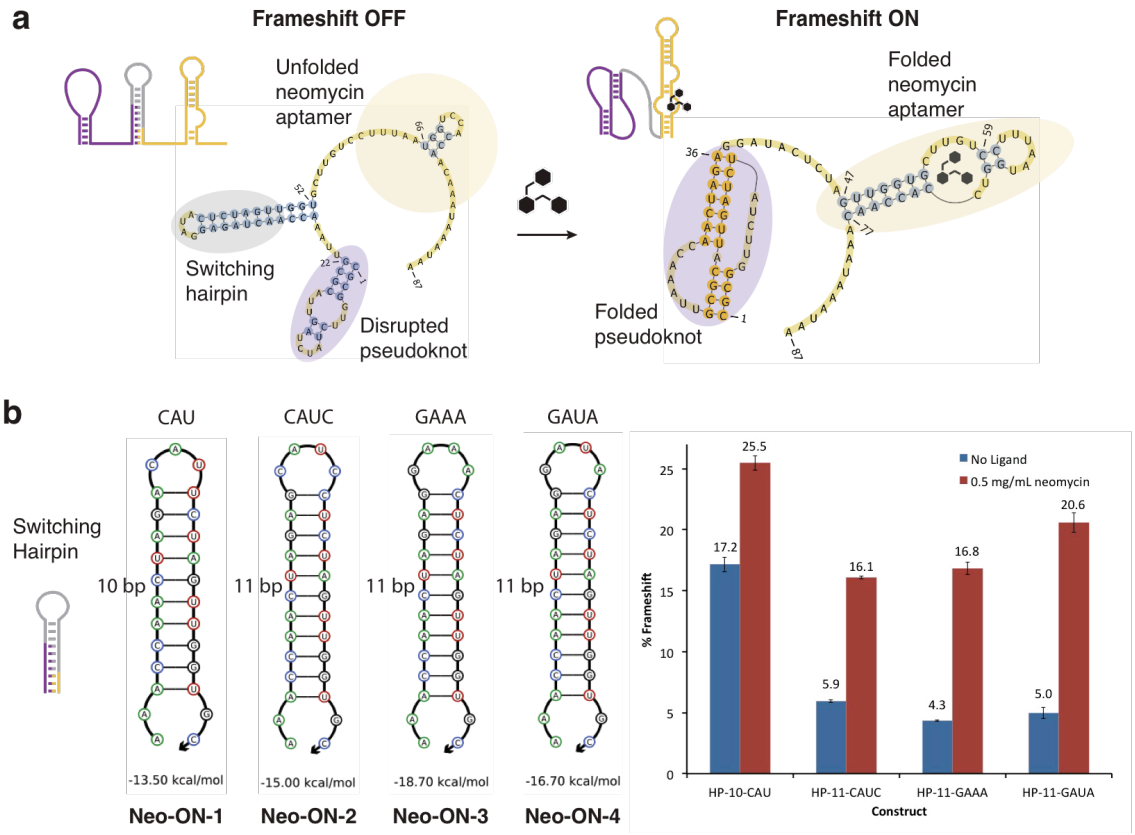


Figure 3-15. Design and characterization of neomycin responsive ON switches. (**a**) General scheme depicting the intended structural rearrangements induced by ligand binding. A representative neomycin responsive ON switch sequence is shown. (**b**) Frameshift levels were tuned by adjusting the length and loop composition of the switching hairpin. In general, basal frameshifting and ligand induced frameshifting is inversely correlated with switching hairpin thermodynamic stability. In this case, by comparison to the theophylline ON switches, hairpin stability was tuned by modifying the tetraloop nucleotides.

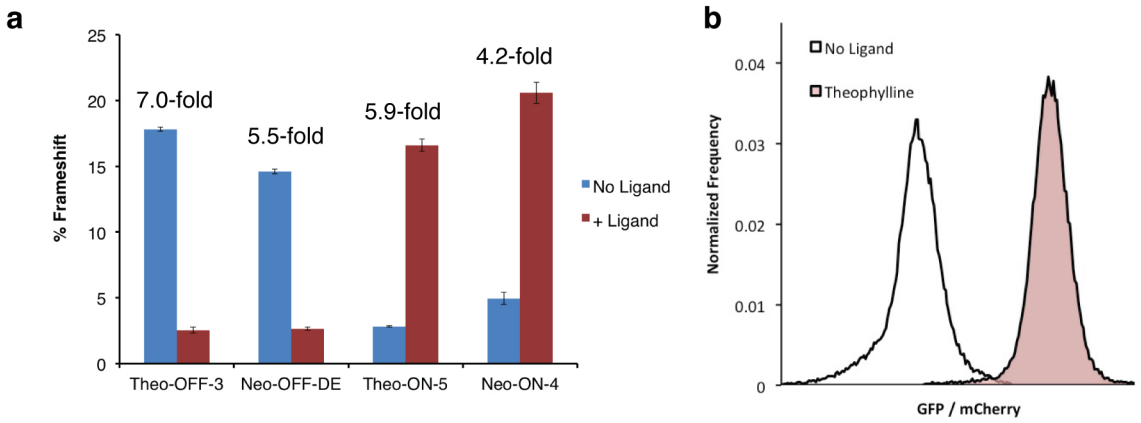


Figure 3-16. Summary of optimized switches. (**a**) The ligand responsiveness of four switch devices as assayed in the dual-FP reporter. Theophylline was used at a concentration of 40 mM; neomycin was used at a concentration of 500 mg/mL. (**b**) Flow cytometry of the Theo-ON-11 switch in the presence and absence of theophylline (40 mM) as visualized by the histogram of the ratio of GFP to mCherry signals (on log scale).

3.2.3 Constructing logic gates and phenotypic controllers with -1 PRF switches

We envisioned exploiting multiple translational reading frames by layering our -1 PRF devices within individual mRNAs to create logic gates and phenotypic controllers. Logic gates are important genetic devices for executing cellular computation and programming biological systems. While diverse logic gate architectures have been reported[19,36,37], most require the expression of multiple components to transduce small-molecule inputs into gene expression output. We recognized the opportunity to construct highly condensed logic gates from a single mRNA transcript using -1 PRF switches. To this end, we constructed both a NOR gate (**Fig. 3-17a**) and an AND gate (**Fig. 3-17b**). Logic gate outputs predicted by individual switch activities are in good agreement with the experimentally obtained results, demonstrating that the frameshift switch devices function independent of one another and their context. Moreover, ON/OFF states are well distinguished by flow cytometry, particularly for the NOR gate. Though not

demonstrated here, it is conceivable that other more complex logic gates could be constructed by layering -1 PRF switches in alternative configurations within mRNAs.

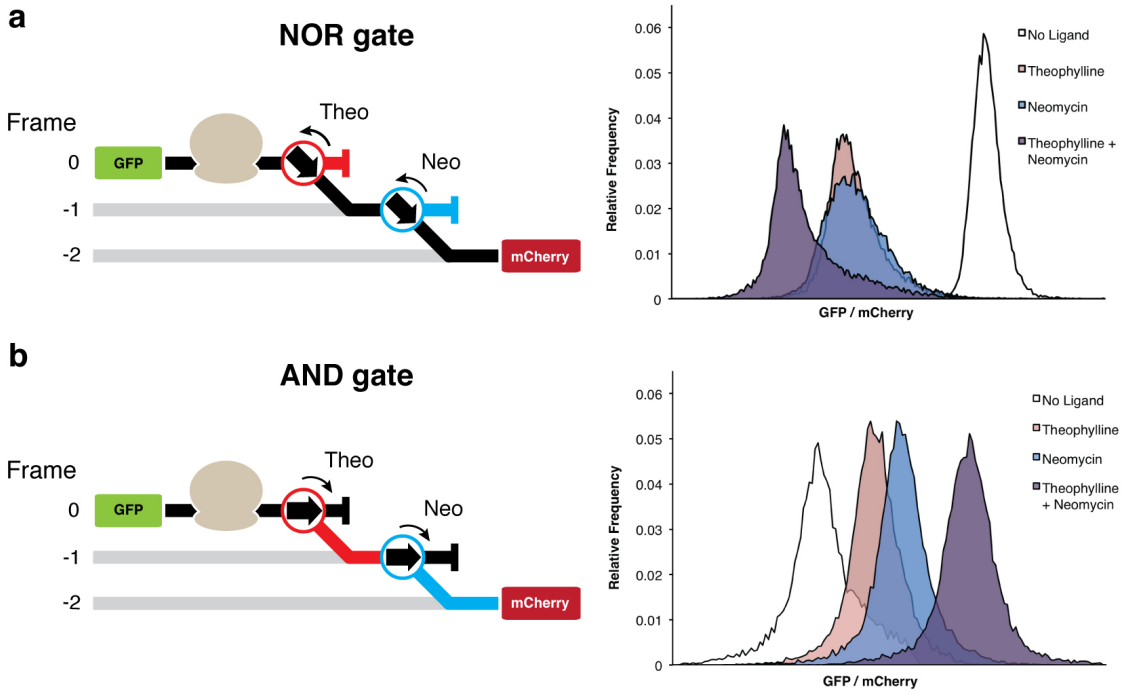


Figure 3-17. Construction of logic gates with layered -1 PRF switch devices. (**a**) The NOR gate is composed the Theo-OFF-3 switch and the Neo-OFF-DE switch layered within a single mRNA. (**b**) The AND gate is composed of the Theo-ON-5 switch and the Neo-ON-4 switch layered within a single mRNA. Translation in the absence of a ligand follows the black path; theophylline directs translation down the red path; neomycin directs translation down the blue path. Activity of the gate was assessed within the dual-FP reporter by flow cytometry. For both gates, mCherry is encoded in the -2 frame with respect to GFP. Theophylline was used at a concentration of 40 mM; neomycin was used at a concentration of 500 mg/mL.

Next, to demonstrate the utility of -1 PRF for controlling cell phenotype and studying interacting proteins *in vivo*, we designed an apoptosis module in yeast using mammalian Bcl-2 family proteins[38]. Though naturally absent from *S. cerevisiae*, heterologously expressed Bcl-2 family members retain their basic functions in yeast[39].

Bax expression in *S. cerevisiae* has been shown to induce cell death through mitochondrial membrane permeabilization[40]. Moreover, co-expression of the pro-survival protein Bcl-xL protects yeast from Bax mediated lethality[39], consistent with its function in mammalian systems.

A third subset of factors termed the BH3-only proteins are proposed to inhibit the pro-survival family members by direct interaction, and perhaps additionally interact with Bax to activate its killing function. Thus, a model has been proposed in which cell fate is dictated by the relative expression of BH3-only and pro-survival proteins in a background of Bax expression[38]. While previous studies in yeast have demonstrated that the BH3-only protein Puma enhances Bax mediated lethality, co-expression of pro-survival Bcl-xL mitigated all lethal effects[41]. Thus, it is unclear to what extent Bcl-xL expression is required to rescue cells from Bax induced cell death and how the relative production of Puma and Bcl-xL influence cell viability.

Our -1 PRF expression system could offer insight into the underlying mechanism of this process by precise ligand-dependent alteration of protein stoichiometry. We designed an expression construct that allows for the simultaneous control of the BH3-only protein Puma by a theophylline responsive ON-switch and the pro-survival protein Bcl-xL by a neomycin responsive OFF-switch (**Fig. 3-18a**). Both proteins are encoded within the same mRNA and are synthesized according to the frameshift activity of their corresponding switch. Importantly, the Puma open reading frame and theophylline responsive ON switch required recoding to remove stop codons from the 0-frame, allowing for non-frameshifting ribosomes to bypass Puma out of frame and access the downstream Neo-OFF switch and Bcl-xL. Using the small molecule ligands, we titrated

the relative levels of Puma and Bcl-xL in a background of Bax expression and observed the effects on cell viability (**Fig. 3-18b**).

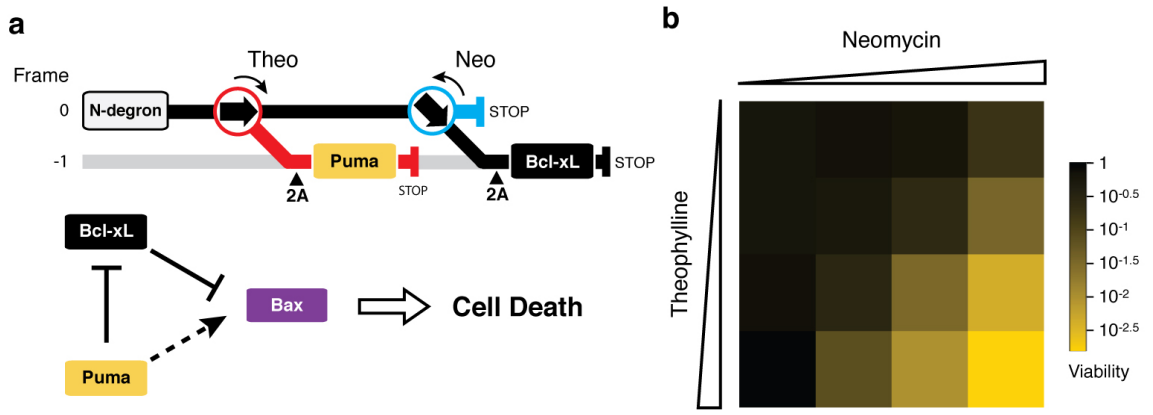


Figure 5. Constructing a cell death module in yeast. (**a**) The apoptosis module construct and biological principle. A single mRNA construct encodes Puma under the control of a theophylline responsive ON-switch and Bcl-xL under the control of a neomycin responsive OFF-switch. 2A peptides encoded N-terminal to the Puma and Bcl-xL open reading frames cleave the functional proteins from the nonsense translation products of the switch devices and alternative reading frames. The latter products are targeted for degradation by an N-degron signal at the N-terminus of the polypeptide. Translation in the absence of a ligand follows the black path, theophylline directs translation down the red path, and neomycin directs translation down the blue path. Relative production of Puma and Bcl-xL controls the ability of Bax to induce cell death. (**b**) Evaluating cell death as a function of relative Puma and Bcl-xL expression. Cells expressing Bax and the Puma/Bcl-xL apoptosis module were grown in various concentrations of neomycin (0 μg/mL, 40 μg/mL, 150 μg/mL, 650 μg/mL) and theophylline (0 mM, 1 mM, 5 mM, 20 mM) and assessed for viability by plating efficiency.

In the absence of ligands, cells were protected from Bax mediated killing through expression of Bcl-xL. Turning down Bcl-xL with neomycin had little effect on cell viability, producing a 4-fold decrease in viability at maximal concentrations. Treatment with theophylline alone to increase Puma production while maintaining high Bcl-xL had

no influence on Bax-mediated killing, consistent with previous reports[41]. However, when cells were treated with both ligands, we observed a cooperative effect building up to a >300-fold decrease in cell viability at maximal concentrations of theophylline and neomycin (**Fig. 3-18b**). These results support the hypothesis that apoptosis is regulated by the relative balance between BH3-only proteins and pro-survival proteins[42], and establish this system as an efficient kill switch in *S. cerevisiae*. Moreover, this represents a functional AND gate with a phenotypically meaningful output.

3.3 Discussion

Our results demonstrate that translation reprogramming can be co-opted and engineered to regulate gene expression in living cells. Because reprogramming occurs through RNA mediated mechanisms, it is amenable to various directed evolution strategies and rational engineering of devices using modular RNA components. Notably, our method applies a combination of these approaches to construct -1 PRF switches. Furthermore, this approach may be applicable to designing -1 PRF switches for higher eukaryote systems, such as mammalian cells, where -1 PRF is commonly observed.

We employed an *in vitro* directed evolution approach that selects for a complex functional phenotype in the setting of a cell lysate. Notably, previous attempts to improve allosteric ribozymes by *in vitro* directed evolution resulted in devices that were non-functional in living cells[18]. This was attributed to potential differences between the *in vitro* and cellular environments. However, our *in vitro* selections were conducted in a cell lysate and enriched for a function that requires direct interaction with the cell's complex biochemical machinery. This, perhaps, improved the likelihood that selection products would retain activity *in vivo*. Analogous to previously reported allosteric selections[17], this

in vitro directed evolution platform could also be adapted for the selection of ligand responsive frameshift devices and the discovery of novel aptamers. Moreover, it could be readily redesigned to select for other forms of translation reprogramming such as +1 frameshifting or stop codon readthrough.

Constitutively active -1 PRF elements are directly applicable for fixed protein expression, and are easily exchanged with individual variants or libraries of variants. The diversity of selection products provides a 150-fold range in frameshifting efficiencies. -1 PRF is also compatible with use in combination with transcription-based expression control tools. The characterization of our selection products by NGS supports future efforts to engineer -1 PRF elements by providing structural information and sequence-activity relationships. Moreover, further analysis of the NGS data could advance our understanding of -1 PRF and support experimental elucidation of the reprogramming mechanism.

Ligand responsive -1 PRF switches provide distinct advantages over other gene-control strategies. Since -1 PRF switches are based on modular RNA parts, it is feasible to scale the number of devices using other aptamer and frameshift components. Whereas engineering proteins that respond functionally to the binding of a ligand remains a challenge, there are better prospects for discovering new RNA aptamers for small molecule recognition. The platform that we established here should be applicable to the construction of other -1 PRF switch devices based on alternative RNA aptamer and frameshift modules to engineer biosensors for molecules of interest.

-1 PRF devices also provide a unique output: stoichiometrically defined products of alternative translation reading frames. This stoichiometric control offers precision and

consistency that can be exploited for various purposes, but especially those that depend on the balance of two functionally interacting proteins. Interestingly, natural biological mechanisms may utilize a similar principle, albeit through transcriptional networks[43]. As demonstrated by our apoptosis module, this form of regulation can be used for achieving robust phenotypic performance in genetically modified organisms.

Importantly, outputs of -1 PRF signals need not be a protein product directly, but may instead be an input frame for another -1 PRF switch device. As a result, -1 PRF signals are very conducive to layering within individual mRNAs. With some creativity and ingenuity, it should be possible to conceive and construct various configurations of -1 PRF signals that can be used to perform complex logic operations and other regulatory functions in the cell. We believe that the tools and methodology developed here will empower our efforts to develop sophisticated genetic programs and advance our capabilities to study and engineer biological systems.

3.4 Experimental Methods

Nucleotide sequences used in this study are detailed in section **3.5**.

Frameshift stimulator library design. The starting library was synthesized as an oligonucleotide (IDT) encoding a modified PreQ1-class I riboswitch[26] with 14 positions randomized in the region surrounding the ligand cavity. This library has a theoretical diversity of 2.68 x 10^8 unique sequences. The selection construct was assembled from 4 oligonucleotides (AVA105, AVA106, AVA107 and AVA108) by PCR with Vent Polymerase (NEB) in a 1-mL reaction. Full-length starting library construct: 5'-TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG ACG ACG ACA AGA CTT TAA ACT AGT TGA CGC G**NN**

NCT A**NN NNN NNN** CGC GTT AAA C**NN N**CT AGA AGG CGG TTC TAT GGG AAT GTC TGG ATT AAG CCA TCA TCA CCA TCA CCA CGG CAG CGG CTA T-3'.

***In vitro* selection for -1 programmed ribosomal frameshifting.** The mRNA display selection protocol described in **Chapter 2** was generally followed. Briefly, the DNA library was *in vitro* transcribed with T7 RNA polymerase (100 uL reaction) and purified by 8.5% Urea-PAGE gel. The resulting RNA was ligated to phospho-A_{27}CC-puromycin (TriLink Biotechnologies) by splint ligation using primers AVA95 and AVA96 as splint oligonucleotides. The puromycin conjugated mRNA templates were purified from non-ligated RNA by 8.5% Urea-PAGE gel. 20 pmoles (1.2×10^{13} molecules) of mRNA display templates were translated at 30˚ C for 1 hour in 100 mL of 40% rabbit reticulocyte lysate (nuclease treated, Promega) supplemented with amino acids, $Mg(OAc)_2$ to 0.5 mM final, and KCl to 100 mM final. After incubation at 30˚ C, the translation reaction was treated with 38 mL of puromycin salt mix and incubated at room temperature for 15 minutes, then placed on ice. The translation reaction was diluted 10-fold with 1.24 mL of chilled FLAG binding buffer. 30 mL of anti-FLAG M2 Affinity Gel (Sigma) was pre-washed twice with FLAG clean buffer and three times with FLAG binding buffer. The diluted translation reaction was incubated with the anti-FLAG affinity gel for 1.5 hrs at 4˚ C with rotation. After incubation, the anti-FLAG gel was separated from the supernatant and washed six times with 0.75-mL portions of chilled FLAG binding buffer. mRNA-peptide fusions were isolated in two elutions by incubating the anti-FLAG gel with 100-mL portions of 200 mg/mL FLAG peptide (Sigma) in FLAG binding buffer for 20 minutes at 4˚ C with rotation. 100 mL of Ni-NTA resin (Qiagen)

was pre-washed three times with ddH_20. The combined FLAG elution fractions were diluted with 200 mL of 2x Ni-NTA binding buffer and incubated with Ni-NTA resin at 4˚ C with rotation for 1 hour. After a series of washes (Supplementary Table XX), mRNA-peptide fusions were eluted with Ni-NTA elution buffer and buffer exchanged into reverse transcription buffer using 30,000-MW cutoff spin concentrators (Amicon). mRNA was reverse transcribed with Superscript II (Invitrogen) for 50 minutes at 42˚ C (Ni-NTA elution, 30 mL; 100 mM RT primer AVA95, 3 mL; 10 mM dNTP mix, 15 mL; RNAse-free ddH_20, 160 mL; 5X First-strand buffer, 60 mL; 0.1M DTT, 30 mL; RNasin, 1 mL, Superscript II Reverse Transcriptase, 1 mL). The reverse transcription reaction was spin-concentrated then amplified with GoTaq polymerase (Promega) using primers AVA108 and AVA109 in a 0.5 mL reaction to generate the DNA for subsequent rounds of selection. Of note, primer AVA109 includes the entire 5' segment of the construct leading up to the first in frame TAG stop codon in order to correct frameshift insertion mutations enriched in the previous round of selection.

Dual-fluorescent protein reporter assay for -1 PRF activity in yeast. The reporter plasmid was constructed by inserting the yEGFP open reading frame (amplified with primer pair AVA111·AVA112) between the GPD promoter and the open reading frame of a yeast-optimized mCherry[44] flanked by a GPD terminator in a pRS425 backbone (high copy, *LEU2* marker). NheI and AatII restriction sites were encoded between the GFP and mCherry open reading frames for replacement of the intervening sequence with a -1 PRF insert. To assay library -1 PRF activity, dsDNA products from the *in vitro* selection were PCR amplified in two steps (primer pairs AVA117·AVA118 and AVA119·AVA118) to add homology to the NheI/AatII digested dual-FP reporter

plasmid. The library was cloned by *in vivo* gap repair into yeast strain Fy251 using high efficiency transformation as previously described[45] and plated on selective medium (synthetic complete agar with leucine dropout containing 2% dextrose; SC-Agar-(Gluc) L-). Individual colonies were isolated and grown in liquid SC-(Gluc) L- media to mid-log phase and the -1 PRF efficiency was determined by comparing quantified GFP and mCherry fluorescence signals to a fusion protein control (calibrated to 100% frameshift). For analysis of the pooled selection products, the rescued transformation was grown in liquid SC-(Gluc) L- selection media for 3 days to enrich for transformants, then analyzed directly by flow cytometry (LSRII).

NGS of *in vitro* selection products. NGS was performed using the Illumina HiSeq platform (Columbia Genome Center). DNA was prepared for sequencing by PCR in two separate batches containing either 5 or 9 variable positions at the start site of the HiSeq read to avoid identical base calls during the initialization phase of sequencing. The *in vitro* selection products were PCR amplified with primer pair AVA317·AVA322 or AVA318·AVA322. Both PCR products were then mixed and amplified with primer pair AVA319·AVA321 to add full HiSeq adaptor sequences. After sequencing, the raw fastq file was trimmed to the scaffold of interest and the copy number of each unique sequence was computed (step 1 of sequence analysis pipeline, see **Appendix A1**).

Analysis of NGS for motif classification. The scaffold and its 14 variable positions define a restricted region of sequence space (2.68×10^8). We constructed a pseudoknot (PK) feature space based on a combination of the nucleotide constraints and our user-defined segment constraints, generating a set of 2,068 individual PK features (**Appendix A1 Fig. A1-2**). We enumerated the full set of unique sequences theoretically

present in the initial library and calculated their compatibilities with the 2,068 PK features. The feature with the greatest amount of base pairing (including G·U) was associated to each sequence. PK compatibilities were similarly calculated for the sequences observed in the *in vitro* selection products. Highly enriched PK features could then be measured by comparing the two distributions. Of the top 10% of PK features by enrichment, those with the highest absolute representation in the selection library defined the broad motif categories. Within a set of PK-compatible sequences, a greedy clustering algorithm was used to divide the set of sequences into the final motifs. The modes of the different motif sets were further characterized in terms of the abundances and entropies of their immediate neighborhood in sequence space. In particular, the modes from the highest abundance motifs were assessed for single- and pairwise- nucleotide variant sensitivity, to reveal potential base pair position and tertiary interactions within a given PK geometry. See **Appendex A-1** for further details

Rational design of -1 PRF switches. Switch constructs were evaluated using RNA secondary structure predictions and thermodynamic calculations (NUPACK and pKiss) as described in **Appendix A2** and **Figure A2-1**. Constructs were assembled from synthesized oligonucleotides (IDT). Switches were cloned into the dual-FP reporter by *in vivo* gap repair of NheI/SalI digested dual-FP reporter plasmid using the -1 PRF switch fragment and an mCherry PCR product containing homology to the switch. Colonies were grown overnight in SC-(Gluc) L- media, then used to seed SC-(Gluc) L- cultures (starting OD_{600} ~ 0.02) with or without the small molecule ligand. Cultures were grown to mid-log phase and measured for -1 PRF activity by quantification of GPF and mCherry fluorescence signals.

Gal4 selection system for *in vivo* directed evolution of -1 PRF switches. The Gal4 selection plasmid was constructed by inserting a minimal Gal4 transcriptional activator between the ADH1 promoter and ADH1 terminator in a pRS425 vector backbone. Gal4 was PCR amplified from yeast genomic DNA in two pieces with primer pairs MS049·AVA155 (Gal4 minimal DNA biding domain, amino acids 1-147) and AVA158·MS050 (Gal4 minimal activation domain, amino acids 768-881). Between the Gal4 binding and activation domains, BamHI and HindIII restriction sites were encoded for replacement of the intervening sequence with a -1 PRF insert. The yeast two-hybrid background strain MaV203 (Invitrogen) was used for all growth assays and growth selections. Control -1 PRF sequences were cloned into the Gal4 selection vector by *in vivo* gap repair of the BamHI/HindIII digested plasmid -1 PRF inserts (two step PCR, first with AVA156·AVA162, then AVA161·AVA162). The Neo-OFF libraries were constructed by PCR assembly of oligonucleotides (Library 1: AVA345·AVA347, then AVA156·AVA346, then AVA161·AVA276; Library 2: AVA345·AVA348, then AVA156·AVA346, then AVA161·AVA276). This generated 2,048 combined theoretical variants. The library was cloned by *in vivo* gap repair into the Gal4 selection plasmid, and the rescued transformation culture was grown in SC-(Gluc) L- at 30 ˚C to enrich for transformants. Positive selection media (SC-(Gluc) HL-, 80 mM 3-AT, 10 mL) was inoculated with the library to a starting OD_{600} = 0.01 and grown for 60 hours at 30 ˚C to OD_{600} = 1.2. SC-(Gluc) L- containing 500 μg/mg neomycin was inoculated with positive selection products and grown from OD_{600} = 0.1 to OD_{600} = 1.5 to allow for protein turnover. This culture was then used to inoculate counterselection media (SC-(Gluc) L-, 500 μg/mg neomycin, 0.05% 5-fluoroorotic acid) to a starting OD_{600} = 0.02 and grown

for 84 hours at 30 ˚C to a final OD_{600} ~ 1.3. DNA was isolated from the counterselection culture and used as PCR template for amplification of the -1 PRF switches and cloning into the dual-FP reporter for assaying frameshift activity (AVA117·AVA357, then AVA119·AVA367, gap repaired with mCherry AVA358·AVA208 PCR product).

Logic gate construction and evaluation. The NOR gate was assembled by fusion PCR of the Theo-OFF-3 switch fragment (AVA119·AVA406) and the Neo-OFF-DE switch fragment (AVA246·AVA247). The AND gate was assembled by fusion PCR of the Theo-ON-5 switch fragment (AVA119·AVA245) and the Neo-ON-4 switch fragment (AVA246·AVA247). Gate inserts were cloned by gap repair of the NheI/SalI digested dual-FP reporter plasmid using the gate PCR fragment and an mCherry PCR product (AVA248·AVA208). Colonies were grown overnight in SC-(Gluc) L- media, then used to seed SC-(Gluc) L- cultures (starting OD_{600} ~ 0.02) with or without the small molecule ligand. Cultures were grown to mid-log phase and analyzed by flow cytometry.

Apoptosis module and cell viability assay. pBM272-3396[46] encoding wild-type mouse Bax under the control of the Gal10 promoter (CEN plasmid, Ura) was obtained from Addgene. The human Puma open reading frame (BBC3, Gene ID: 27113) nucleotide sequence lacking the ATG start codon was codon optimized for yeast expression, manually recoded to remove stop codons from its +1 frame, and ordered synthesized (IDT). The nucleotide sequence for human Bcl-xL (BCL2L1, Gene ID: 598) was codon optimized for yeast expression and ordered synthesized (IDT). The N-degron signal (Ubiquitin-Arg-LacI(1-37)) was amplified from an existing plasmid with primer pair AVA462·AVA364. The full construct (**Table 3-3**) was assembled by conventional molecular biology techniques and chromosomally integrated at the LEU2 locus in an

Fy251 background strain giving AA01 (*leu2Δ::LEU2-pTDH3-[Apoptosis Module]*). This strain was then transformed with pBM272-3396, giving strain AA02. For control viability assays, cultures of AA02 were grown for 72 hours at 30 ˚C without Bax induction in synthetic complete media with uracil dropout containing 2% raffinose (SC-(Raf) U-) and either no ligand, theophylline (20 mM), neomycin (650 μg/mL), or both theophylline (20 mM) and neomycin (650 μg/mL). For assessing viability with Bax expression, cultures of AA02 were grown in synthetic complete media with uracil dropout containing 2% raffinose and 2% galactose (SC-(Gal) U-) with varying concentrations of theophylline (0 mM, 1 mM, 5 mM, or 20 mM) and neomycin (0 μg/mL, 40 μg/mL, 150 μg/mL, or 650 μg/mL) for 72 hours at 30 ˚C. After this time, the OD of each culture was measured, diluted to a standard cell density, and then plated at the appropriate dilution in triplicate on SC-Agar-(Gluc) U- plates. Colonies from each plate for the same condition were counted and averaged (**Table 3-5**).

3.5 DNA sequences and apoptosis viability

Table 3-1. Sequences of -1 PRF OFF-switches.

Switch	Sequence
Theo-OFF-1	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCAAC**TAG**AGATACCAGCATCGTCTGATGCCCTTGGCAGC **TCTA**AAAAAATAAATAATAAAATTAAA
Theo-OFF-2	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCAA**CTAGAGATACCAGCATCGTCTGATGCCCTTGGCAGC TCTAG**AAAAATAAATAATAAAATTAAA
Theo-OFF-3	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCA**ACTAGAGATACCAGCATCGTCTGATGCCCTTGGCAGC TCTAGT**AAAATAAATAATAAAATTAAA
Neo-OFF-1	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCAAC**TAGAGGCTTGTCCTTTAATGGTCCCTCTA**AAAAACA TAAATAATAAAATTAAA
Neo-OFF-2	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCAA**CTAGAGGCTTGTCCTTTAATGGTCCCTCTAG**AAAAC A TAAATAATAAAATTAAA
Neo-OFF-3	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCA**ACTAGAGGCTTGTCCTTTAATGGTCCCTCTAGT**AAACA TAAATAATAAAATTAAA
Neo-OFF-4	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCA**ACTAGAAGCTTGTCCTTTAATGGTCCTTCTAGT**AAACA TAAATAATAAAATTAAA
Neo-OFF-DE	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAA CCAA**CTAGTTGCTTGTCCTTTAATGGTCCAACTAG**AAAACA TAAATAATAAAATTAAA

Sequences begin with the slippery site and end with the insulating sequence. The pseudoknot sequence is underlined and the aptamer sequence bolded.

Table 3-2. Sequences of -1 PRF ON-switches.

Switch Name	**Sequence (5' to 3')**
Theo-ON-1	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGACATACTA**GTTGGT**GATACCAGCATCGTCTTGA* **TGCCCTTGGCAGCACC**CAAAATAAATAATAAAATTAAA
Theo-ON-2	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGACATTCTAGTT**GGT**GATACCAGCATCGTCTTGA* **TGCCCTTGGCAGCACC**CAAAATAAATAATAAAATTAAA
Theo-ON-3	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGCATCCTCTA**GTTGGT**GATACCAGCATCGTCT* **TGATGCCCTTGGCAGCACC**CAAAATAAATAATAAAATTAAA
Theo-ON-4	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGCATCCTCTAGT**TGGT**GATACCAGCATCGTCT* **TGATGCCCTTGGCAGCACCA**CAAATAAATAATAAAATTAAA
Theo-ON-5	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGCATCCTCTAG**TTGGT**GATACCAGCATCGTCT* **TGATGCCCTTGGCAGCACCAA**AAATAAATAATAAAATTAAA
Theo-ON-6	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGCATCCTCTA**GTTGGT**GATACCAGCATCGTCT* **TGATGCCCTTGGCAGCACCAAC**AATAAATAATAAAATTAAA
Neo-ON-1	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGACATTCTA**GTTGGT**GCTTGTCCTTTAATGGTCC* **ACCAAC**AAATAAATAATAAAATTAAA
Neo-ON-2	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGCATCCTCTA**GTTGGT**GCTTGTCCTTTAATGG* **TCCACCAAC**AAATAAATAATAAAATTAAA
Neo-ON-3	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGGAAACTCTA**GTTGGT**GCTTGTCCTTTAATGG* **TCCACCAAC**AAATAAATAATAAAATTAAA
Neo-ON-4	TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAA*A CCAACTAGAGGATACTCTA**GTTGGT**GCTTGTCCTTTAATGG* **TCCACCAAC**AAATAAATAATAAAATTAAA

Sequences begin with the slippery site and end with the insulating sequence. The pseudoknot sequence is underlined, the switching hairpin is italicized, and the aptamer sequence bolded.

Table 3-3. Sequences of logic gates and the apoptosis module.

Construct	Sequence (5' to 3')
NOR gate	**TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAACCAAC TAGAGATACCAGCATCGTCTGATGCCCTTGGCAGCTCTAG**AAAAAT AAATAATAAAATTAAAATGGGTTCAGGTGAACAATCAAAGAC**TTTAA ACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAACCAACTAGTT GCTTGTCCTTTAATGGTCCAACTAG**AAAACATAAATAATAAAATTAA A
AND gate	**TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAACCAAC TAGAGCATCCTCTAGTTGGTGATACCAGCATCGTCTTGATGCCCTT GGCAGCACCAA**AAATAAATAATAAAATTAAAATGGGTTCAGGTGAAC AATCAAAGAC**TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCG TTAAACCAACTAGAGGATACTCTAGTTGGTGCTTGTCCTTTAATGG TCCACCAAC**AAATAAATAATAAAATTAAA
Apoptosis module open reading frame	ATGCAGATTTTCGTCAAGACTTTGACCGGTAAAACCATAACATTGGAA GTTGAATCTTCCGATACCATCGACAACGTTAAGTCGAAAATTCAAGAC AAGGAAGGTATCCCTCCAGATCAACAAAGATTGATCTTTGCCGGTAAG CAGCTAGAAGACGGTAGAACGCTGTCTGATTACAACATTCAGAAGGAG TCCACCTTACATCTTGTGCTAAGGCTAAGAGGTGGTAGGCATGGATCC GGAGCTTGGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTG GCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTA ATGCAGGGCAGCCTGCAGGGCGACTACAAGGACGACGACGACAAGAC **TTTAAACTGGTTGACGCGGTTCGTATCTTGTTACGCGTTAAACCAA CAAGAGAATCTCTTGTTGGTGATACCAGCATCGAAAGATGCCCTTG GCAGCACCAA**AATAATAATAAAATTAAAAATGTCGACGGTTCTGGTG GTTCTGGTGAAGGTAGAGGTTCTTTGCTTACTTGTGGTGATGTCGAAGA GAATCCTGGT*CCGAAGTTTGGTATGGGTTCTGCTCAAGCTTGTCCATG TCAAGTCCCAAGAGCTGCTTCCACTACCTGGGTCCCATGTCAAATTTGT GGTCCTCAACCATCTTTATCCTTGGCCGAACAACACTTGGAATCCCCAG TTCCATCCGCTCCTGGTGCCTTGGCCGGTGGTCCTACCCAAGCTGCTCC AGGTGTTAGAGGTGAAGAAGAACAATGGGCTAGAGAAATCGGTGCCC AATTGCGTAGAATGGCCGATGATCTTAACGCTCAATACGAAAGAAGAA GACAAGAAGAACAACAAAGACATAGACCATCCCCTTGGAGAGTTTTAT ATAACCTTATTATGGGTTTGTTGCCATTACCTAGAGGTCACCGTGCTCC AGAAATGGAACCAAATTAAGCTAGCAGGTTCAGGTGAACAATCAAAG AC**TTTAAACTAGTTGACGCGGTTCTATCTAGTTACGCGTTAAACCA ACTAGTTGCTTGTCCTTTAATGGTCCAACTAG**AAAACATAAATAAT AAAATTAAAGACGTCGGTTCTGGTGGTTCTGGTGAAGGTAGAGGTTCA TTGTTAACATGTGGAGATGTTGAGGAAAATCCTGGT*CCGATGTCCCAA TCTAACAGAGAATTAGTCGTTGACTTCTTGTCTTACAAATTGTCTCAAA AGGGTTACTCTTGGTCCCAATTTTCCGATGTCGAAGAAAATAGAACTG AAGCTCCAGAAGGTACCGAGTCTGAAATGGAAACTCCATCCGCTATCA

	ACGGTAACCCATCCTGGCACTTGGCTGATTCTCCAGCTGTCAACGGTGC CACTGGTCACTCCTCCTCCTTAGATGCTCGTGAGGTTATTCCAATGGCC GCCGTCAAGCAAGCTTTGAGAGAAGCTGGTGATGAATTCGAATTGAGA TACAGAAGAGCCTTCTCTGACTTGACCTCTCAATTGCATATCACTCCAG GTACTGCTTACCAATCCTTCGAACAAGTTGTTAACGAATTGTTCAGAGA CGGTGTTAACTGGGGTAGAATTGTCGCTTTTTTCTCTTTCGGTGGTGCC TTATGTGTTGAATCTGTTGACAAGGAAATGCAAGTCTTGGTTTCCAGAA TTGCTGCTTGGATGGCTACCTACTTGAATGACCACTTGGAACCATGGAT TCAAGAAAACGGTGGTTGGGATACTTTCGTCGAGTTGTACGGTAATAA CGCTGCCGCTGAATCTAGAAAGGGTCAAGAAAGATTCAATCGTTGGTT CTTGACTGGTATGACTGTCGCTGGTGTTGTCTTGTTGGGTTCCTTATTCT CCAGAAAGTGA

-1 PRF switches are bolded and underlined. * indicates position of 2A peptide cleavage.

Table 3-4. Sequences of oligonucleotides used in this work.

Name	Sequence (5' to 3')
AVA-95	TTT TTT TTT TTT ATA GCC GCT GCC
AVA105	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG
AVA106	CGC GTC AAC TAG TTT AAA GTC TTG TCG TCG TCG TCC TTG TAG TCC ATT GTA ATT GTA AAT
AVA107	GAC AAG ACT TTA AAC TAG TTG ACG CG**N NN**C TA**N NNN NNN N**CG CGT TAA AC**N NN**C TAG AAG GCG GTT CTA TGG GAA TGT C
AVA108	ATA GCC GCT GCC GTG GTG ATG GTG ATG ATG GCT TAA TCC AGA CAT TCC CAT AGA ACC GCC
AVA109	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG ACT ACA AGG ACG ACG ACG ACA AGA CTT TAA ACT AG
AVA111	ATA AAC ACA CAT AAA CAA ACA AAG AAT TCA TGT CTA AAG GTG AAG AAT TAT TCA CTG GTG
AVA112	GCT AGC TTT GTA CAA TTC ATC CAT ACC ATG GGT AAT ACC
AVA117	AGC TAG CGG CAG CGG CGA CTA CAA GGA CGA CGA CGA CAA GAC
AVA118	CAT ATT ATC TTC TTC ACC TTT TGA AAC CAT GAC GTC TCC AGA CAT TCC CAT AGA ACC GCC
AVA119	ATT ACC CAT GGT ATG GAT GAA TTG TAC AAA GCT AGC GGC AGC GGC GAC TAC
AVA155	GTA GTG CGC CAG AAC CAC TGC CGG ATC CCG ATA CAG TCA ACT GTC TTT GAC
AVA156	CGG CAG TGG TTC TGG CGC ACT ACA AGG ACG ACG ACA CAA GAC TTT AAA CTA
AVA158	TAT GGG AAT GTC TGG AAA GCT TGC CAA TTT TAA TCA AAG TGG GAA TAT TGC
AVA161	TAA CAA AGG TCA AAG ACA GTT GAC TGT ATC GGG ATC CGG CAG TGG TTC TGG CGC ACT AC
AVA162	AGC AAT ATT CCC ACT TTG ATT AAA ATT GGC AAG CTT TCC AGA CAT TCC CAT AGA ACC GCC
AVA208	CAC ACA GGA AAC AGC TAT GAC CAT G
AVA245	AGT CTT TGA TTG TTC ACC TGA ACC CAT TTT AAT TTT ATT ATT TAT TTT TGG TGC TGC C
AVA246	ATG GGT TCA GGT GAA CAA TCA AAG ACT TTA AAC TAG TTG ACG CGG TTC
AVA247	ATA GCC ATA TTA TCT TCT TCA CCT TTT GAA ACC
AVA248	ATG GTT TCA AAA GGT GAA GAA GAT AAT ATG GC
AVA276	AGC AAT ATT CCC ACT TTG ATT AAA ATT GGC AAG CTT AGC CAT ATT ATC TTC TTC AC
AVA317	TTC CCT ACA CGA CGC TCT TCC GAT CTN NNN NGA CAA GAC TTT AAA CTA GTT GAC GCG

AVA318	TCC CTA CAC GAC GCT CTT CCG ATC TNN NNN NNN NGA CAA GAC TTT AAA CTA GTT GAC GCG
AVA319	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC
AVA321	CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CAG ACG TGT GCT C
AVA322	GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CAC ATT CCC ATA GAA CCG CCT TCT AG
AVA345	GAC ACA AGA CTT TAA ACT AGT TGA CGC GGT TCT ATC TAG TTA CGC GTT AAA CCA ACT AG
AVA346	CAA GCT TAG CCA TAT TAT CTT CTT CAC CTT TTG AAA CCA TTT TAA TTT TAT TAT TTA TGT
AVA347	CCT TTT GAA ACC ATT TTA ATT TTA TTA TTT ATG TTT NCT AGN NGG ACC ATT AAA GGA CAA GCN NCT AGT TGG TTT AAC GCG TAA CTA GAT
AVA348	CAC CTT TTG AAA CCA TTT TAA TTT TAT TAT TTA TGT NNN CTA GNG GAC CAT TAA AGG ACA TCN CTA GTT GGT TTA ACG CGT AAC TAG AT
AVA357	CTT AGC CAT ATT ATC TTC TTC ACC TTT TGA AAC C
AVA358	GGT TTC AAA AGG TGA AGA AGA TAA TAT GGC
AVA364	CCT TGT AGT CGC CCT GCA GGC TGC CCT GCA TTA ATG AAT C
AVA367	TTA AAA ATG TCG ACG GTT CTG GTC AAT TGC TTA ACT TCG ATT TAC TTA AAT TGG
AVA406	GTC TTT GAT TGT TCA CCT GAA CCC ATT TTA ATT TTA TTA TTT ATT TTT CTA GAG CTG CC
AVA462	CAA AAT GCA GAT TTT CGT CAA GAC TTT GAC CGG
MS049	GAT ATC GAC GTC ATG AAG CTA CTG TCT TCT ATC GAA C
MS050	GTA TGC GCT AGC TTA CTC TTT TTT TGG GTT TGG TGG GG

Table 3-5. Colony counts[a] for viability assay of apoptosis module.

Conditions[b]	Plate 1	Plate 2	Plate 3	Average	Stand. Err.
T-0 N-0	64	70	62	65.3	2.4
T-1 N-0	77	52	68	65.7	7.3
T-2 N-0	90	78	67	78.3	6.6
T-3 N-0	102	114	100	105.3	4.4
T-0 N-1	82	78	72	77.3	2.9
T-1 N-1	63	51	55	56.3	4.3
T-2 N-1	24	34	31	29.7	3.6
T-3 N-1	10	7	7	8.0	1.0
T-0 N-2[c]	62.7	65	61.5	63.1	1.0
T-1 N-2[c]	24.5	26.2	29.5	26.7	1.8
T-2 N-2[c]	2.5	4.1	3.8	3.5	0.5
T-3 N-2[c]	1.3	1.2	0.6	1.0	0.2
T-0 N-3[c]	21.8	18.3	16.8	19.0	1.5
T-1 N-3[c]	3.3	3.5	4.4	3.7	0.3
T-2 N-3[c]	0.3	0.8	0.3	0.5	0.2
T-3 N-3[c]	0.3	0.1	0.1	0.2	0.1
T-0 N-0[d]	75	72	80	75.7	2.3
T-3 N-0[d]	99	101	94	98	2.1
T-0 N-3[d]	90	103	110	101	5.9
T-3 N-3[d]	127	129	139	131.7	3.7

[a]Cultures were diluted to a standard cell density based on OD_{600} and so that approximately 10^3 cells were plated on SC(gluc) L- selection. Colony counts are reported as number of colonies per 10^3 cells plated.

[b]Ligand concentrations. T-0: No theophylline; T-1: 1 mM theophylline; T-2: 5 mM theophylline; T-3: 20 mM theophylline. N-0: No neomycin; N-1: 40 μg/mL neomycin; N-2: 150 μg/mL neomycin; N-3: 650 μg/mL neomycin.

[c]Cultures were diluted 10-fold less in this case so that 10^4 cells were plated.

[d]Cultures in which Bax expression was not induced with galactose, grown with raffinose as the sole carbon source.

3.6 References

1. Zaher, H. S. & Green, R. Fidelity at the Molecular Level: Lessons from Protein Synthesis. *Cell* **136,** 746–762 (2009).
2. Gesteland, R. F. & Atkins, J. F. Recoding: Dynamic Reprogramming of Translation. *Annu. Rev. Biochem.* **65,** 741–768 (1996).
3. Firth, A. E. & Brierley, I. Non-canonical translation in RNA viruses. *J. Gen. Virol.* **93,** 1385–1409 (2012).
4. *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. (Springer New York, 2010). at <http://link.springer.com/10.1007/978-0-387-89382-2>
5. Mountford, P. S. & Smith, A. G. Internal ribosome entry sites and dicistronic RNAs in mammalian transgenesis. *Trends Genet.* **11,** 179–184 (1995).
6. Felipe, P. de, Hughes, L. E., Ryan, M. D. & Brown, J. D. Co-translational, Intraribosomal Cleavage of Polypeptides by the Foot-and-mouth Disease Virus 2A Peptide. *J. Biol. Chem.* **278,** 11441–11448 (2003).
7. Liu, C. C. & Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **79,** 413–444 (2010).
8. Isaacs, F. J., Dwyer, D. J. & Collins, J. J. RNA synthetic biology. *Nat. Biotechnol.* **24,** 545–554 (2006).
9. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).
10. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249,** 505–510 (1990).
11. Robertson, D. L. & Joyce, G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344,** 467–468 (1990).
12. Jenison, R. D., Gill, S. C., Pardi, A. & Polisky, B. High-resolution molecular discrimination by RNA. *Science* **263,** 1425–1429 (1994).
13. Berens, C., Thain, A. & Schroeder, R. A tetracycline-binding RNA aptamer. *Bioorg. Med. Chem.* **9,** 2549–2556 (2001).
14. Tang, J. & Breaker, R. R. Rational design of allosteric ribozymes. *Chem. Biol.* **4,** 453–459 (1997).

15. Klauser, B., Atanasov, J., Siewert, L. K. & Hartig, J. S. Ribozyme-Based Aminoglycoside Switches of Gene Expression Engineered by Genetic Selection in S. cerevisiae. *ACS Synth. Biol.* **4,** 516–525 (2015).
16. Townshend, B., Kennedy, A. B., Xiang, J. S. & Smolke, C. D. High-throughput cellular RNA device engineering. *Nat. Methods* **advance online publication,** (2015).
17. Koizumi, M., Soukup, G. A., Kerr, J. N. & Breaker, R. R. Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nat. Struct. Biol.* **6,** 1062–1071 (1999).
18. Link, K. H. *et al.* Engineering high-speed allosteric hammerhead ribozymes. *Biol. Chem.* **388,** 779–786 (2007).
19. Win, M. N. & Smolke, C. D. Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science* **322,** 456–460 (2008).
20. Ausländer, S. *et al.* A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat. Methods* **11,** 1154–1160 (2014).
21. Chen, Y. Y., Jensen, M. C. & Smolke, C. D. Genetic control of mammalian T-cell proliferation with synthetic RNA regulatory systems. *Proc. Natl. Acad. Sci.* **107,** 8531–8536 (2010).
22. Galloway, K. E., Franco, E. & Smolke, C. D. Dynamically Reshaping Signaling Networks to Program Cell Fate via Genetic Controllers. *Science* **341,** 1235005 (2013).
23. Brierley, I. Ribosomal frameshifting on viral RNAs. *J. Gen. Virol.* **76,** 1885–1892 (1995).
24. Jacks, T., Madhani, H. D., Masiarz, F. R. & Varmus, H. E. Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. *Cell* **55,** 447–458 (1988).
25. Ivanov, I. P., Gesteland, R. F. & Atkins, J. F. Antizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit. *Nucleic Acids Res.* **28,** 3185–3196 (2000).
26. Yu, C.-H., Luo, J., Iwata-Reuyl, D. & Olsthoorn, R. C. L. Exploiting preQ1 Riboswitches To Regulate Ribosomal Frameshifting. *ACS Chem. Biol.* **8,** 733–740 (2013).

27. Hsu, H.-T., Lin, Y.-H. & Chang, K.-Y. Synergetic regulation of translational reading-frame switch by ligand-responsive RNAs in mammalian cells. *Nucleic Acids Res.* **42,** 14070–14082 (2014).

28. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci.* **94,** 12297–12302 (1997).

29. Liu, R., Barrick, J. E., Szostak, J. W. & Roberts, R. W. Optimized synthesis of RNA-protein fusions for in vitro protein selection. *Methods Enzymol.* **318,** 268–293 (2000).

30. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. & Szostak, J. W. Constructing high complexity synthetic libraries of long ORFs using In Vitro selection. *J. Mol. Biol.* **297,** 309–319 (2000).

31. Janssen, S. & Giegerich, R. The RNA shapes studio. *Bioinformatics* **31,** 423–425 (2015).

32. Weigand, J. E. *et al.* Screening for engineered neomycin riboswitches that control translation initiation. *RNA* **14,** 89–97 (2008).

33. Fields, S. & Song, O. A novel genetic system to detect protein protein interactions. *Nature* **340,** 245–246 (1989).

34. Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E. & Boeke, J. D. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc. Natl. Acad. Sci.* **93,** 10315–10320 (1996).

35. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32,** 170–173 (2011).

36. Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P. & Endy, D. Amplifying Genetic Logic Gates. *Science* **340,** 599–603 (2013).

37. Moon, T. S., Lou, C., Tamsir, A., Stanton, B. C. & Voigt, C. A. Genetic programs constructed from layered logic gates in single cells. *Nature* **491,** 249–253 (2012).

38. Youle, R. J. & Strasser, A. The BCL-2 protein family: opposing activities that mediate cell death. *Nat. Rev. Mol. Cell Biol.* **9,** 47–59 (2008).

39. Sato, T. *et al.* Interactions among members of the Bcl-2 protein family analyzed with a yeast two-hybrid system. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 9238–9242 (1994).

40. Priault, M., Camougrand, N., Kinnally, K. W., Vallette, F. M. & Manon, S. Yeast as a tool to study Bax/mitochondrial interactions in cell death. *FEMS Yeast Res.* **4,** 15–27 (2003).

41. Gallenne, T. *et al.* Bax activation by the BH3-only protein Puma promotes cell dependence on antiapoptotic Bcl-2 family members. *J. Cell Biol.* **185,** 279–290 (2009).

42. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. *Nat. Rev. Mol. Cell Biol.* **15,** 49–63 (2014).

43. Lee, R. E. C., Walker, S. R., Savery, K., Frank, D. A. & Gaudet, S. Fold Change of Nuclear NF-κB Determines TNF-Induced Transcription in Single Cells. *Mol. Cell* **53,** 867–879 (2014).

44. Keppler-Ross, S., Noffz, C. & Dean, N. A New Purple Fluorescent Color Marker for Genetic Studies in Saccharomyces cerevisiae and Candida albicans. *Genetics* **179,** 705–710 (2008).

45. Pirakitikulr, N., Ostrov, N., Peralta-Yahya, P. & Cornish, V. W. PCRless library mutagenesis via oligonucleotide recombination in yeast. *Protein Sci. Publ. Protein Soc.* **19,** 2336–2346 (2010).

46. Gross, A. *et al.* Biochemical and genetic analysis of the mitochondrial response of yeast to BAX and BCL-X(L). *Mol. Cell. Biol.* **20,** 3125–3136 (2000).

Chapter 4

Versatile Diaryl Ether Intermediates for the Gram-Scale Synthesis of Oxazine and Xanthene Fluorophores

---

*The content of this chapter has been published in:

A.V. Anzalone, T.Y Wang, Z. Chen, V.W. Cornish "A Common Diaryl Ether Intermediate for the Gram-Scale Synthesis of Oxazine and Xanthene Fluorophores." *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 650-4.

4.0 Chapter Outlook

Fluorescence based methods for visualizing the localization of molecules within cells are critical for studying basic biological processes. In many ways, the fluorescent proteins (FPs) have revolutionized cell biology, offering user-friendly fluorescent tags that achieve exquisite labeling specificity through genetic encoding. While FPs play an important role in basic biological imaging, they are currently limited to labeling protein molecules that can tolerate an FP fusion. Moreover, recently developed imaging modalities, especially those that rely on single molecule detection as in super-resolution imaging, benefit from fluorescent tags with higher photon outputs and photostability. As a result, synthetic small molecule fluorophores, which often have superior photophysical qualities, have become increasingly important reagents for these more demanding imaging technologies. Nonetheless, methodologies for synthesizing fluorophores are often troubled by poor yields, numerous byproducts, and challenging product isolation. This makes further functionalization of the core scaffolds challenging and hinders the development of new fluorescent probes. In this chapter, we address these synthetic difficulties and present a general approach for the preparation of oxazine and xanthene fluorophores via a common diaryl ether intermediate. First, we identified a suitable copper(I) based catalyst system for the efficient preparation of electron rich diaryl ethers. Second, we developed a scheme for efficiently converting diaryl ethers into oxazine fluorophores by a two-step process that results in high yields and nearly no side products. Lastly, we developed a scheme to convert diaryl ethers to rosamine fluorophores by a tandem Friedel-Crafts acylation and cyclization process. Notably, an example from each class was performed on the gram-scale.

4.1 Introduction

Recent advances in fluorescence spectroscopy have driven the demand for small molecule dyes with improved photophysical and fluorescence properties[1–3]. In addition to their development as live-cell protein labeling reagents[4,5], engineered fluorescent dyes have also been developed as environmental sensors that can provide readouts of local viscosity, pH, solute concentration, and electrical potential[6–10]. Current and future developments in this field, especially those relevant to single-molecule and cellular imaging, depend on the synthesis of customized fluorescent dyes that emit in the red region of the visible spectrum, have high extinction coefficients and quantum yields, and display high photostability[11,12].

Amongst our best dyes for these purposes are those from the oxazine and xanthene classes, exemplified by commercial compounds ATTO-655 and Alexa Fluor-594, respectively[11,13,14]. Derivitization of these parent scaffolds by chemical modification of commercially available dyes is often limited, either due to cumbersome functionalization of the parent compounds or prohibitive costs in obtaining sufficient quantities of dye for carrying out the necessary synthetic steps. Thus, the *de novo* synthesis of fluorescent dyes from basic organic building blocks is an essential aspect of technology development. Despite the obvious importance of these molecules and the evident need for improved synthetic methodologies, oxazines and xanthenes are still largely synthesized using methods reported decades or more ago that do not take advantage of the efficiencies of modern chemical transformations[15–17].

Aiming to develop improved technologies for *in vivo* super resolution imaging, we sought to prepare derivatives of the commonly utilized oxazine ATTO-655[14]. Due to

the expense associated with obtaining large quantities of this dye and the lack of commercial availability of other desirable analogs, we set out to synthesize the oxazines following previously described methods[16,17]. These prior works rely on coupling a pair of aminophenol derivatives, one of which is substituted with an electrophilic nitroso or diazo functionality, by heating the two components in an acidic medium (**Fig. 4-1**, prior works).

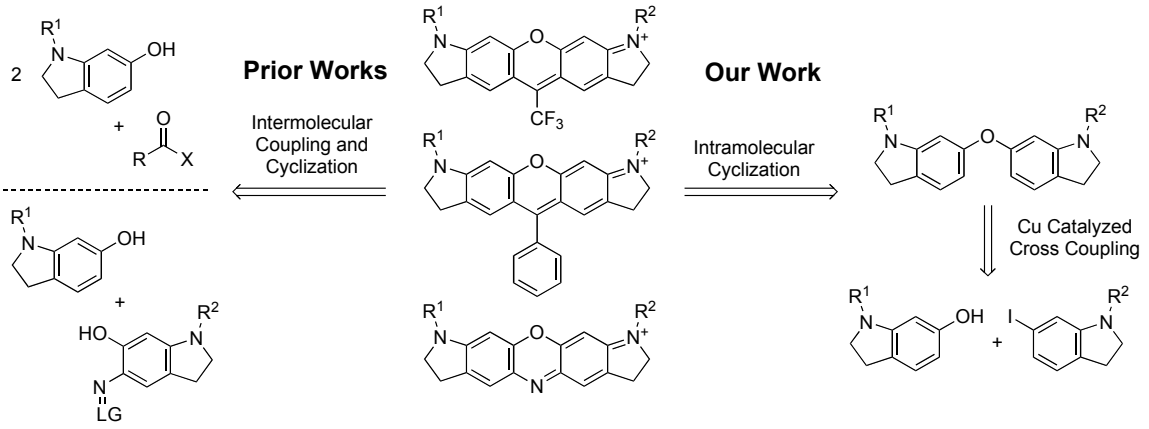


Figure 4-1. Retrosynthetic analysis for oxazine and xanthene fluorophores. Prior works utilized multicomponent couplings, whereas our approach pre-assembles the diaryl scaffold for facile and versatile conversion to various fluorophores. LG = leaving group. X = H or RCO_2–.

Though the aminophenol intermediates are readily prepared, the final coupling and cyclization reactions frequently result in low yields (~15%) and necessitate the use of preparative scale reverse phase high performance liquid chromatography (prep-HPLC) in order to obtain material of satisfactory purity. In our hands, the quantities obtained by this method proved insufficient to reasonably carry out the remaining steps in the synthesis of our final targets, and we decided that scaling up the process to obtain the desired

quantities was impractical. As a result, we concluded that the existing approach was not a viable synthetic route and chose to pursue alternative synthetic strategies.

Within the oxazine core structure**,** we identified a previously unexplored diaryl-ether disconnection, which is exploited in the retrosynthetic analysis shown in **Figure 4-1**. We envisioned that a step-wise coupling and cyclization process would proceed in higher overall yield when compared to the classical coupling strategy and, more importantly, would yield fewer byproducts in the final step. Thus, isolation and purification could be carried out using standard flash chromatography techniques amenable to gram-scale operations. In pursuit of this synthetic strategy, it was necessary to examine methods for synthesizing the diaryl ether in a versatile and robust fashion.

Here, we report a novel and scalable synthetic approach to the assembly of the widely used oxazine and xanthene fluorophores via a diaryl-ether intermediate (**Fig. 4-1**). Taking advantage of recent developments in transition metal catalysis, we constructed electronically activated diaryl ethers to serve as tethered nucleophiles, reacting with a range of substrates to undergo cyclization to cationic fluorescent compounds (**Fig. 4-1**, our work). Final products were provided in good overall yields, were amenable to purification using standard silica-based normal phase flash chromatography, and, significantly, could be prepared on the gram scale.

4.2 Results

4.2.1 Assembly of diaryl ethers by Cu(I)-mediated coupling chemistry

In addition to the more traditional dihydroquinoline and tetrahydroquinoline derivatives, we also chose to explore an indoline-based scaffold to study this synthetic sequence. The precursor indoles provide a large number of building blocks from which to

start, and we were also interested in the photophysical properties of the resulting oxazine products, which are not well described for the indoline class. Preparation of the indoline coupling partners proved to be straightforward by Gribble reduction and alkylation of the commercially available indole starting materials (**Fig. 4-2**)[18].

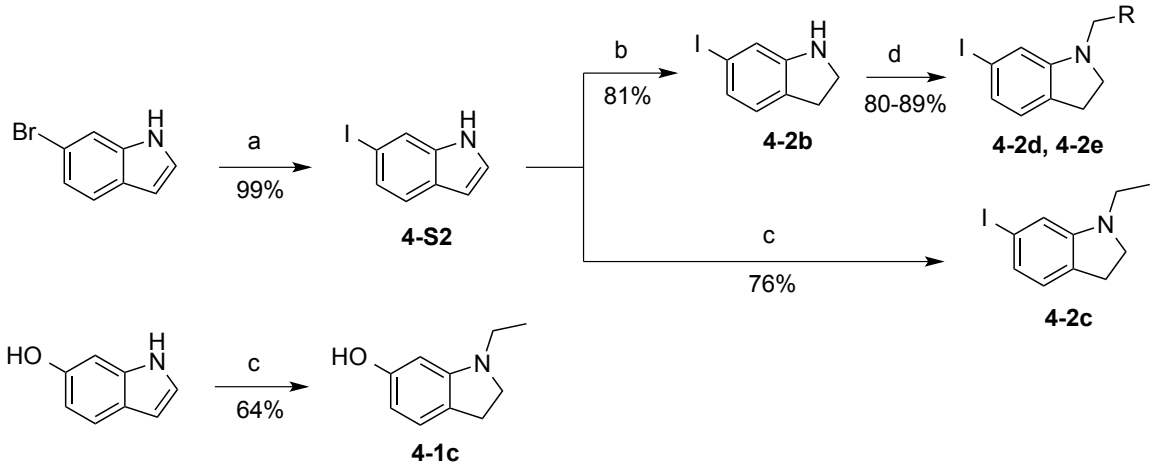


Figure 4-2. Preparation of hydroxyindolines and iodoindolines for coupling reactions. Conditions: (a) NaI, CuI (5-10 mol %), N,N'-dimethylethylenediamine (10-20 mol %), 1,4-dioxane, 110 ^{o}C, 22 hrs. (b) $NaBH_3CN$, AcOH, RT. (c) $NaBH_4$, AcOH, RT. (d) Alkyl halide, base, heat (see **4-2d** and **4-2e**).

The Ullmann ether synthesis poses a potential route to the diaryl ether structural motif by copper(I) promoted reaction between phenols and aryl halides[19]. Although attractive in theory, the classical reaction conditions typically employ strong base, stoichiometric quantities of copper, and heating at temperatures in excess of 200 ˚C. Additionally, the reaction is known to be highly substrate dependent and, when successful, commonly results in only modest yields. To overcome these limitations and eliminate the harsh reaction conditions, we chose to explore a more contemporary method that utilizes palladium catalysis[20]. However, the coupling between phenol **1c** and

its corresponding triflate **4-S8** unexpectedly proceeded predominantly in a Heck-type manner from reaction at C-5 of the indoline, producing the biaryl phenol derivative **4-S9** as the major product (**Fig. 4-3**). Thus, the palladium catalyzed coupling reaction does not appear to be a useful method for synthesizing diaryl ethers when electron rich carbon nucleophiles, such as those that exist in our system, are present.

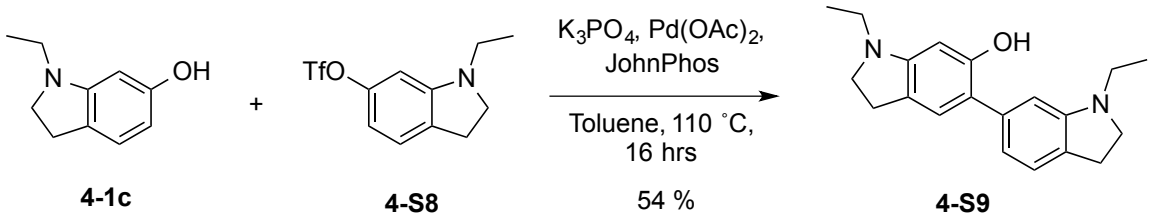
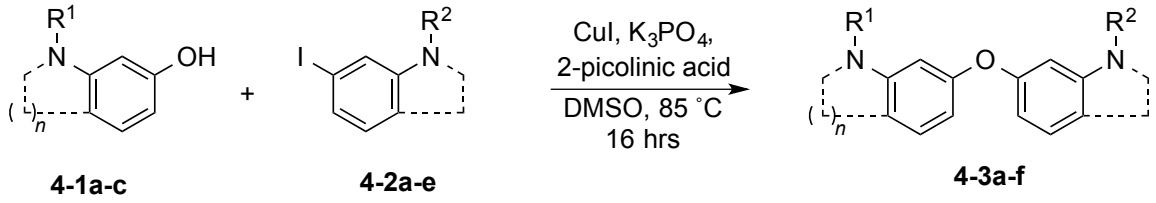


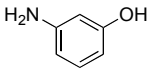
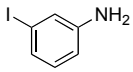
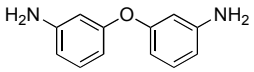
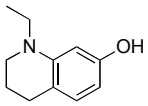
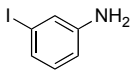
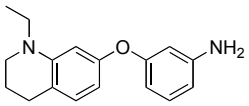
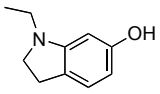
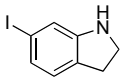
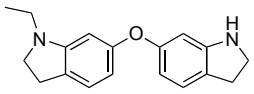
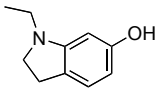
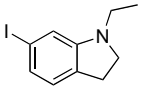
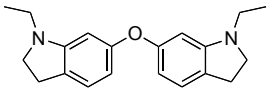
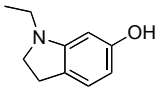
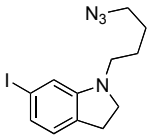
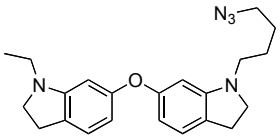
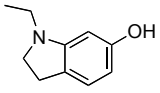
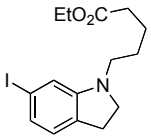
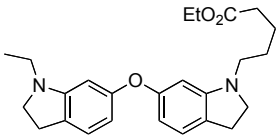
Figure 4-3. Palladium catalyzed coupling of electron rich phenols and aryl triflates. Conditions: 1.2 eq. phenol, 2 mol% $Pd(OAc)_2$, 3 mol% 2-(di-tert-butylphosphino)biphenyl, 2 eq. K_3PO_4, toluene, 110 ˚C.

We next turned to recent developments in ligand-assisted copper catalyzed coupling reactions[21]. Based on work reported by Buchwald and coworkers, we found that the coupling between phenol **4-1b** and 3-iodoaniline (**4-2a**) provided the diaryl ether **4-3b** in 82% yield when carried out in the presence of catalytic copper iodide (10 mol %) and 2-picolinic acid (20 mol %) at 85 ˚C in DMSO (**Table 4-1**, entry 2). This catalytic system offers the advantage of orthogonality with the aniline functional group. As opposed to aryl iodides, aryl bromides were found to be very poor substrates in our system, with nearly no product formation observed even at elevated temperatures (180 ˚C) and increased catalyst loadings (50 mol %). Nonetheless, aryl iodides were readily prepared from the corresponding commercially available aryl bromides by employing a copper promoted halogen exchange reaction[22].

Table 4-1. Diaryl ethers by copper(I) catalyzed couplings of phenols and aryl iodides.[a]



4-1a-c **4-2a-e** **4-3a-f**

Entry	Phenol	Aryl Iodide	Diaryl Ether	% Yield
1	 **4-1a**	 **4-2a**	 **4-3a**	90
2	 **4-1b**	 **4-2a**	 **4-3b**	82
3	 **4-1c**	 **4-2b**	 **4-3c**	88
4	 **4-1c**	 **4-2c**	 **4-3d**	80
5	 **4-1c**	 **4-2d**	 **4-3e**	77
6	 **4-1c**	 **4-2e**	 **4-3f**	80

[a] Conditions: 1.2 eq. phenol, 10 mol % copper(I) iodide, 20 mol % 2-picolinic acid, 2.0 eq. K_3PO_4, DMSO, 85 ^oC, 24 hrs.

Couplings carried out according to the aforementioned conditions provided a range of substituted diaryl ethers in good yields (**Table 4-1**). Following these coupling reactions, anilines **4-3a** and **4-3b** were transformed to their dihydroquinoline derivatives via a modified Skraup reaction (**Fig. 4-4**) and subsequently alkylated and reduced where appropriate.

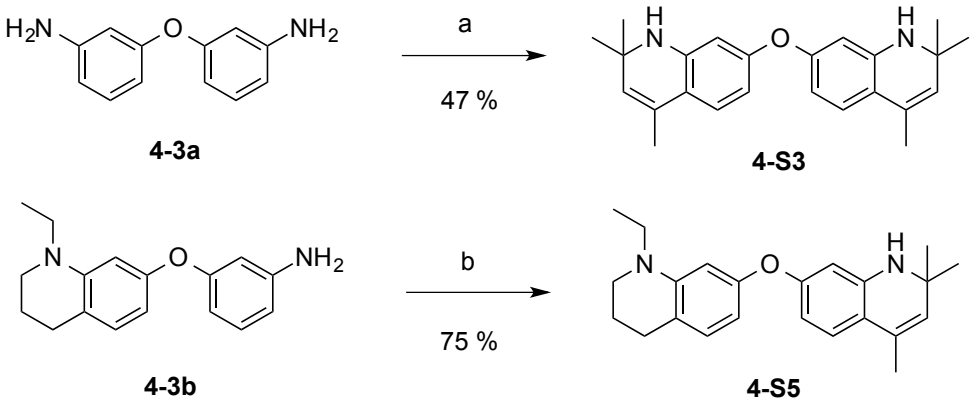


Figure 4-4. Modified Skraup reactions to prepare dihydroquinolines. Conditions: (a) 0.25 eq. TsOH, acetone, RT, 48 hrs. (b) 0.5 eq. TsoH, 2,2-dimethoxypropane, RT, 5 days.

4.2.2 Synthesis of oxazine fluorophores

Having established a reliable means of synthesizing the critical diaryl ether intermediates, we next explored conditions for converting these compounds to their corresponding oxazine dyes. This transformation was readily accomplished by reaction of diaryl ether **4-3d** with one equivalent of 4-nitrobenzene diazonium tetrafluoroborate, followed by heating the corresponding diazene **4-4d** with *p*-toluenesulfonic acid (TsOH) to 65 ˚C in ethanol (**Fig. 4-5**). The latter process proceeded with near-quantitative conversion and following trivial silica based flash chromatography to remove the liberated *p*-nitroaniline, we obtained the oxazine-tosylate salt in high purity and 90% yield.

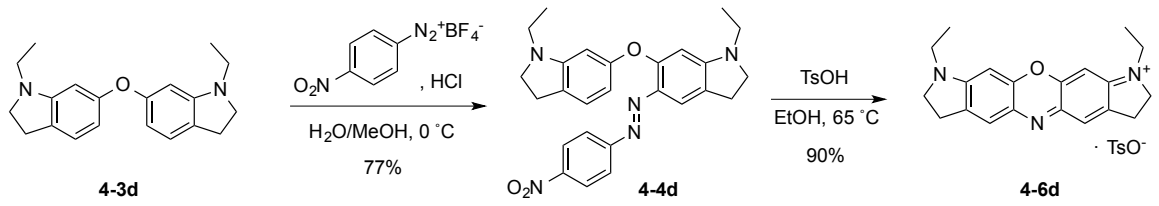
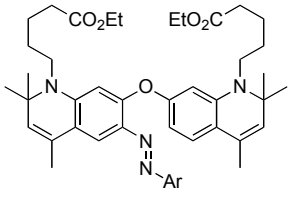
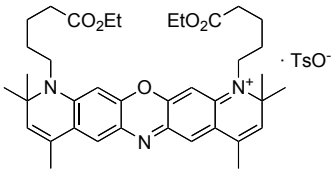
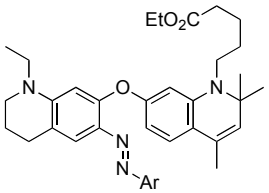
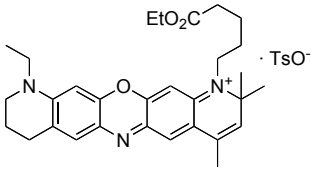


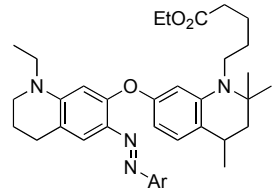
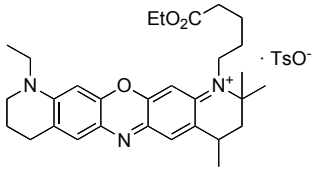
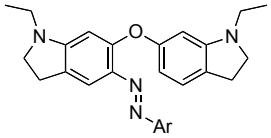
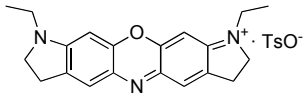
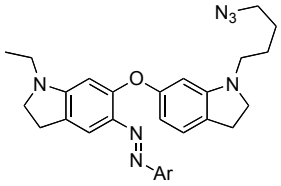
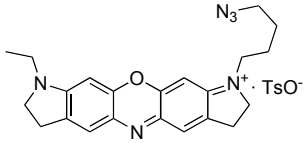
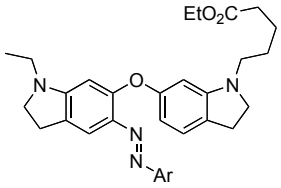
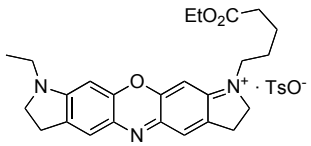
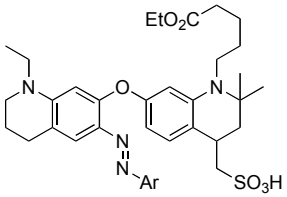
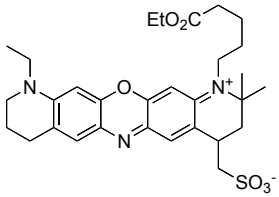
Figure 4-5. Reaction sequence for preparing oxazines from diaryl ethers.

Through implementing this strategy, several substituted oxazine dyes with various spectroscopic properties were synthesized in good to excellent yields from the corresponding diazene compounds (**Table 4-2**). Of note, over one gram of oxazine **4-6f** was prepared by this method in 94% yield, demonstrating the scalability of our approach (**Table 4-2**, entry 6). Additionally, **4-6g**, the sulfonated analog of **4-6c** resembling commercially available ATTO-655 was synthesized in good yield by this method.

Table 4-2. Synthesis of substituted oxazine dyes.[a]

Diazene **4-4a-f** —TsOH, EtOH, 65 ˚C→ Oxazine · Tosylate Salt **4-6a-f**

Entry	Diazene [b,c]	Oxazine	% Yield
1	 **4-4a**	 **4-6a**	74
2	 **4-4b/4-5b**	 **4-6b**	85

3	 **4-4c/4-5c**	 **4-6c**	87
4	 **4-4d**	 **4-6d**	90
5	 **4-4e/4-5e**	 **4-6e**	78
6[d]	 **4-4f/4-5f**	 **4-6f**	94
7	 **4-4g/4-5g**	 **4-6g**	83

[a] Conditions: 3.0 eq TsOH, EtOH, 65 ˚C, 4-8 hrs. [b] Ar = *p*-NO_2Ph. [c] Diazenes were obtained and used as regioisomeric mixtures, which cyclize to a single oxazine product. [d] Reaction conducted on the gram-scale.

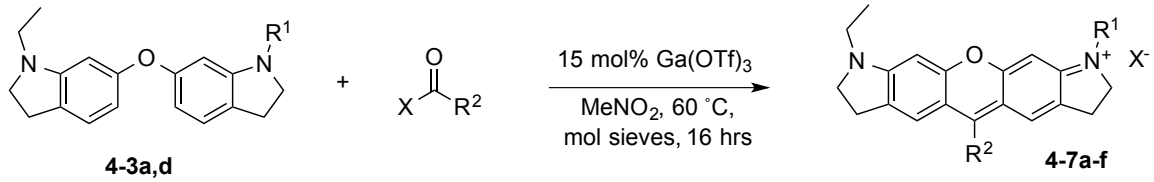
4.2.3 Synthesis of rosamine fluorophores by tandem Friedel-Crafts acylation/cyclization

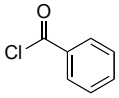
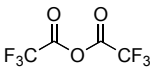
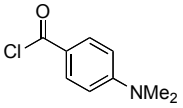
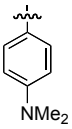
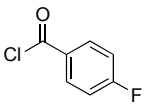
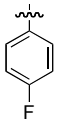
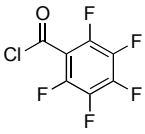
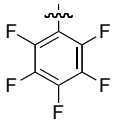
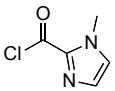
To further expand the scope of this methodology, we examined other popular fluorescent dyes for similar diaryl ether disconnections and identified xanthene dyes, exemplified by rhodamine and rosamine, as potential targets (**Fig. 4-1**). The classical

method for synthesizing these aminoxanthenes calls for heating two equivalents of an aminophenol with one equivalent of a carboxylic acid anhydride or aldehyde under acidic conditions ($ZnCl_2$ or H_2SO_4) often at high temperatures (**Fig. 4-1**, prior works)[15]. When this strategy is employed with derivatized analogs, the reactions typically result in modest to poor yields and make isolation of the product difficult by conventional purification methods[23]. We sought to prepare these fluorescent dyes from our diaryl ethers in an analogous manner to the synthesis of the oxazine dyes. After screening several Lewis acid catalysts and reaction conditions, we found that $Ga(OTf)_3$ catalyzed the tandem Friedel-Crafts acylation-cyclization reaction between the aforementioned diaryl ethers and aromatic acid chlorides to provide the corresponding xanthene dyes in modest to good yields with substantial recoverable starting material (**Table 4-3**)[24–26].

Supplementing the reaction with additional $Ga(OTf)_3$ did not lead to further product formation, suggesting that a byproduct formed in the course of the reaction was inhibiting forward progression. Possibly, this was due to substrate inactivation by HCl, which is generated in the course of the Friedel-Crafts acylation and by hydrolysis of the acid chloride with water generated in the cyclization step. Interestingly, in the case of the p-Me_2N derivative (**Table 4-3**, entry 3), nearly full consumption of starting material was observed, suggesting that the dimethylamino group may be capable of buffering the reaction to allow for increased conversion. Higher conversions of starting material could be obtained by re-subjecting the crude product mixture (following workup) to the initial reaction conditions.

Table 4-3. Tandem catalytic Friedel-Crafts acylation/cylization reaction for the synthesis of xanthene fluorophores.[a]



Entry	R^1		R^2	Product	% Yield (B.R.S.M)[b]
1	Et			**4-7a**	56 (92)
2[c, e]				**4-7b**	83
3[d]	Et			**4-7c**	77 (82)
4	Et			**4-7d**	48 (89)
5	Et			**4-7e**	53 (93)
6[d]	Et			**4-7f**	36 (81)

[a] Conditions: 8 eq acid chloride, 15 mol % $Ga(OTf)_3$, $MeNO_2$, 60˚C, 4Å mol sieves, 16 hrs. [b] B.R.S.M = % yield based on recovered starting material. [c] No Lewis acid catalyst required, 2.2 eq of TFAA in DCM at RT for 12 hrs. [d] 5 eq of the acid chloride was used. [e] Reaction conducted on the gram-scale.

Of note, this synthetic strategy allows for the synthesis of asymmetrically functionalized xanthene dyes, which are not accessible by the classical coupling strategy. Additionally, several dyes were prepared in significantly higher yields when compared to traditional syntheses. For example, compound **4-7b**, an asymmetric xanthene, was synthesized in 83% yield simply by reaction of the diaryl ether with 2.2 equivalents of trifluoroacetic anhydride at room temperature without a Lewis acid catalyst. By comparison, rhodamine 700, a similar dye possessing the trifluoromethyl substituent at the *meso* carbon, is reported to be prepared in only 5% yield from aminophenols[27]. Although this current work is limited to the synthesis of rosamines with fully substituted anilines, it may be possible to synthesize mono- or non-alkylated analogs by minor modification of this synthetic approach.

4.2.4 Photophysical characterization of oxazine and rosamine fluorophores

Figure 4-6 shows the absorption and fluorescence spectra for several of the oxazine and xanthene dyes synthesized. Notably, the increased conjugation of oxazines **4-6a** and **4-6b** results in significant red-shifts in the absorption and fluorescence spectra. Similarly, xanthenes **4-7b**, **4-7e** and **4-7f** possessing electron withdrawing side chains also display red-shifted spectra. The dimethylamino-substituted rosamine **4-7c** displays pH dependent fluorescence, with a near 30-fold increase in fluorescence intensity in acidic conditions (data not shown). Spectral properties of the dyes in aqueous solution are provided in **Table 4-4**.

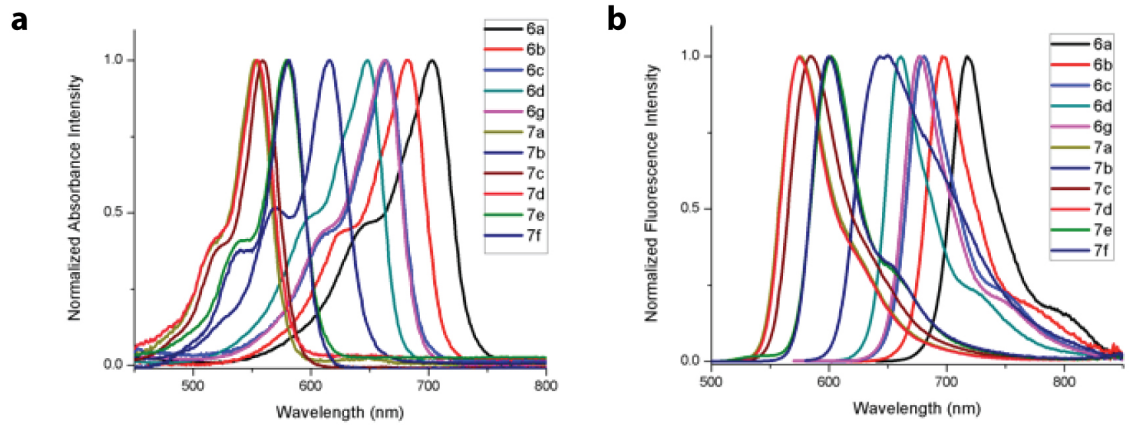


Figure 4-6. Spectral characterization of synthesized fluorophores. (a) Absorbance and (b) fluorescence spectra of various oxazine and xanthene derivatives. Spectra were obtained in H_2O, with the exception of **4-7c**, which was obtained in aqueous 50 mM HCl.

Table 4-4. Spectral properties of fluorescent dyes in H_2O.[a]

Dye	$l_{max\ abs}$ (nm)	e_{max} ($M^{-1}cm^{-1}$)	l_{fluor} (nm)	Fwhm[b] (nm)	F_f
4-6a	703	50,000	717	42	0.08
4-6b	682	69,000	696	44	0.09
4-6c	664	67,000	681	45	0.24
4-6d	648	66,000	661	43	0.11
4-6g	663	97,000	677	43	0.20
4-7a	553	58,000	576	49	0.19
4-7b	616	61,000	643	90	0.08
4-7c	559	79,000	585	53	0.14
4-7d	555	60,000	575	47	0.21
4-7e	579	68,000	602	44	0.27
4-7f	581	57,000	600	42	0.24

[a] Measurements were made in H_2O, with the exception of **4-7c**, which was measured in aqueous 50 mM HCl. [b] Full-width at half-maximum height.

4.3 Discussion

In summary, we have established a high-yielding, scalable synthetic route for the preparation of the widely used oxazine and xanthene based fluorescent dyes using diaryl ether intermediates. Compared to previously existing synthetic methodologies, our work provides a versatile means for preparing these fluorescent dyes and eliminates the need for tedious and expensive purifications. Following this synthetic approach, a number of oxazine and xanthene fluorophores were synthesized and characterized. With proper synthetic planning, we believe this to be a general and widely applicable approach to the synthesis of derivatives of oxazine and xanthene dyes, and may facilitate the development of novel fluorophores and probes with unique properties.

4.4 Experimental Methods

General Experimental Methods

All reactions were carried out under an argon atmosphere. Anhydrous 1,4-dioxane, dimethylsulfoxide, tetrahydrofuran, acetonitrile and dichloromethane were purchased as Sure Seal™ bottles from Sigma-Aldridge. Potassium phosphate tribasic was purchased from Sigma-Aldrich and ground to a fine powder just prior to use. Sodium iodide powder was purchased from Alfa Aesar and ground to a fine powder just prior to use. Copper(I) iodide (99.999%) powder was purchased from Sigma-Aldrich. 6-hydroxyindole and 6-bromoindole were purchased from Chem-Impex International. Gallium(III) trifluoromethanesulfonate was obtained from Acros. All other reagents were commercially available and used without further purification. Flash chromatography was performed using a Teledyne ISCO CombiFlash R_f system. NMR spectra were recorded

on Bruker DRX-300 and DRX-400 instruments. The following abbreviations were used to describe multiplicities: s = singlet; d = doublet; t = triplet; q = quartet; m = multiplet; br = broad. High-resolution mass spectrometry (HRMS) was performed by the Columbia University Mass Spectroscopy Core Facility with a JOEL HX110 mass spectrometer by means of fast atom bombardment (FAB). Absorbance measurements and spectra were obtained with a Tecan Infinite 200 and fluorescence measurements and spectra were obtained with a Horiba Scientific Fluorolog-3 spectrofluorometer. Measurements were performed in quartz cuvettes with a 1-cm path length using solutions with absorbance under 0.1 to prevent inner filter and other non-linear effects.

General procedure for the Copper(I) promoted coupling of phenols and aryl iodides.

According to the published protocol by Maiti and Buchwald[21], an oven dried, sealable glass vessel was charged with a magnetic stirbar, the phenol (2.40 mmol), potassium phosphate (4.00 mmol, 849 mg), copper(I) iodide (0.20 mmol, 38 mg), 2-picolinic acid (0.40 mmol, 49 mg), and the aryl iodide, if a solid (2.00 mmol). The vessel was then fitted with a rubber septum, evacuated under vacuum and backfilled with argon. This process was repeated 3 times. The vessel was then charged with DMSO (4.0 mL), or if the aryl iodide is a liquid, the vessel was charged with the aryl iodide as a solution in DMSO. The rubber septum was removed and the reaction vessel was immediately sealed tightly with a Teflon screw cap. The reaction was then heated to 85 ˚C for 16-24 hours. After cooling to room temperature, the reaction was diluted with 10 mL of water and extracted with ethyl acetate (25 mL, 4x). The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a crude residue. Purification

by flash chromatography (hexanes/ethyl acetate) afforded the diaryl ethers as colorless oils, which were stored at -20 ˚C under inert atmosphere (these compounds become colored upon prolonged exposure to air at room temperature or storage as a solution in a halogenated solvent, such as chloroform).

General procedure for the reaction of diarly ethers with p-nitrobenzenediazonium tetrafluoroborate.

The diarly ether (1.00 mmol) was dissolved in methanol or ethanol (2 mL), cooled to 0 ˚C in an ice bath, then treated with aqueous 2N HCl (5 mL). After sufficient time to cool, the reaction was treated with *p*-nitrobenzenediazonium tetrafluoroborate (1.00 mmol) in 5 mg portions over 5 minutes, then stirred at 0 ˚C for 1 hour. The reaction was then diluted with dichloromethane (10 mL), quenched by the slow addition of saturated aq. $NaHCO_3$ and extracted with dichloromethane (3 x 20 mL). The combined organic layers were dried over Na_2SO_4 and concentrated in vacuo to a deep red-purple residue. Flash chromatography (10-30% ethyl acetate/hexanes) provided the diazene as a mixture of regioisomers in the case of asymmetrical diaryl ether starting materials (both of which transform to the same oxazine product upon cyclization). Full characterization of diazene **4-4d** is provided.

General procedure for cyclization of diazene-diaryl ethers to oxazines

The diazene diaryl ether (0.2 mmol) was dissolved in anhydrous ethanol (20 mL) and treated with p-toluenesulfonic acid monohydrate (0.6 mmol). The deep red reaction mixture was heated to 65 ˚C – 70 ˚C, becoming deep blue after a short period of heating.

The reaction was monitored by TLC (8% MeOH/DCM) and continued until complete conversion of starting material to the product (typically ~4-8 hrs). After cooling to room temperature, the reaction was treated with 4% $NaHCO_3$ and extracted with dichloromethane (3 x 25 mL). The combined organics were dried over anhydrous Na_2SO_4, and then concentrated to a crude blue-red residue. Products were purified by flash chromatography (silica gel) using a gradient of 0-10% MeOH/DCM.

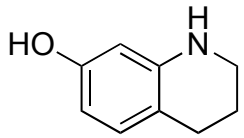
General procedure for the $Ga(OTf)_3$ catalyzed tandem Friedel-Crafts acylation/cyclization.

A small roundbottom flask or sealable vial was purged with argon and charged with a magnetic stirbar, activated 4 Å molecular sieves, the diaryl ether (0.50 mmol), anhydrous nitromethane (0.80 mL), the acid chloride (4.0 mmol for **4-7a**, **4-7d**, **4-7e**; 2.5 mmol for **4-7c**, **4-7f**), and gallium triflate (0.075 mmol). The reaction was heated to 60 ˚C for 16 hours. After cooling to RT, the reaction was diluted with dichloromethane (25 mL) and 4% aq. $NaHCO_3$ (30 mL) and extracted with dichloromethane (3 x 25 mL). The combined organic layers were dried over anhydrous Na_2SO_4 and then concentrated to crude darkly colored residues. Products were purified by flash chromatography (silica gel) using a gradient of 0-10% MeOH/DCM.

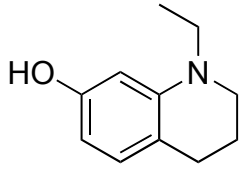
Dye Characterization:

All dyes were HPLC purified (C18 semi-prep, MeCN/H_2O) prior to characterization. Measurements were performed in DI water with the exception of **4-7c**, which required 50 mM HCl for fluorescence. Molar extinction coefficients were determined using the Beer-

Lambert law by measuring the absorbance of known concentrations of dye solution. Fluorescence quantum yields were determined using the comparative method[28,29]. In brief, fluorescence emission was compared to that of a quantum yield reference solution under the same excitation and collection conditions.

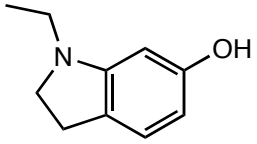


1,2,3,4-tetrahydroquinolin-7-ol (4-S1). 7-hydroxy-3,4-dihydroquinolin-2(1*H*)-one (4.90 g, 30.0 mmol) in a slurry of THF (125 mL) was cooled in an ice bath and treated with lithium aluminum hydride (1.82 g, 1.60 mmol). After complete addition, the reaction was heated to reflux overnight. Upon cooling to room temperature, the reaction was quenched by the addition of aqueous saturated ammonium chloride (50 mL) and extracted with ethyl acetate (3 x 40 mL). The combined organic layers were washed with brine and dried over anhydrous Na_2SO_4, then concentrated to provide **4-S1** in 4.34 g as an orange solid (98%). This compound has been previously characterized[17]. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.81 (d, *J* = 8.1 Hz, 1H), 6.12 (dd, *J* = 8.1, 2.5 Hz, 1H), 6.01 (d, *J* = 2.4 Hz, 1H), 3.46 (s, 1H), 3.29 (t, *J* = 5.7, 2H), 2.70 (t, *J* = 6.4 Hz, 2H), 1.93 (td, *J* = 11.4, 6.3 Hz, 2H).

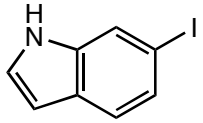


1-ethyl-1,2,3,4-tetrahydroquinolin-7-ol (4-1b). A solution of **4-S1** (1.00 g, 6.70 mmol) in glacial acetic acid (25 mL) was treated with sodium borohydride (1.01 g, 26.8 mmol). After stirring at room temperature for 2 hours, a TLC of the reaction indicated approximately 50% consumption of starting material. At this time, acetaldehyde was added in small portions (reaction TLC'd after each addition) until consumption of starting material was achieved. The reaction was then concentrated under reduced pressure to a thick residue, which was diluted with 60 mL of saturated aqueous $NaHCO_3$ and neutralized with solid $NaHCO_3$. The resulting solution was extracted with ethyl acetate (3 x 35 mL). The combined organic layers were washed with brine (30 mL) and dried over anhydrous Na_2SO_4, then concentrated to a crude orange oil. Purification by flash chromatography (5-20% ethyl acetate/hexanes) provided 1.09 g (92%) of **4-1b** as a light

orange crystalline solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.82 (d, *J* = 8.0 Hz, 1H), 6.24 (s, 1H), 6.12 (d, *J* = 8.0 Hz, 1H), 3.33 (q, *J* = 7.1 Hz, 2H), 3.28 (t, *J* = 5.7 Hz, 2H), 2.70 (t, *J* = 6.4 Hz, 2H), 2.02 – 1.91 (m, 2H), 1.18 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 155.41, 146.33, 130.27, 115.73, 103.21, 98.89, 48.64, 45.92, 27.80, 22.84, 10.94. HRMS (FAB) Calcd for $C_{11}H_{15}NO^+$ $[M]^+$: 177.1154; found 177.1162.

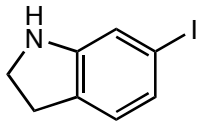


1-ethylindolin-6-ol (4-1c). 6-hydroxyindole (2.66 g, 20.0 mmol) was dissolved in glacial acetic acid (65 mL) and treated with sodium borohydride (3.78 g, 100 mmol) in small portions at room temperature. After 4 hours, the reaction was concentrated in vacuo to a thick residue, which was then diluted with 60 mL of saturated aqueous $NaHCO_3$ and neutralized with solid $NaHCO_3$. The resulting solution was extracted with ethyl acetate (35 mL, 3x). The combined organic layers were washed with brine (30 mL) and dried over anhydrous Na_2SO_4, then concentrated in vacuo and dried under high vacuum to provide an orange crude solid. Purification by flash chromatography (5-20% ethyl acetate/hexanes) provided 2.09 g (64%) of **4-1c** as a tan solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.90 (d, *J* = 7.8 Hz, 1H), 6.10 (dd, *J* = 7.8, 2.1 Hz, 1H), 6.02 (d, *J* = 2.1 Hz, 1H), 4.87 (br s, 1H), 3.37 (t, *J* = 8.2 Hz, 2H), 3.12 (q, *J* = 7.2 Hz, 2H), 2.90 (t, *J* = 8.2 Hz, 2H), 1.20 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 156.28, 153.81, 125.12, 122.92, 104.95, 96.93, 53.32, 43.78, 28.14, 11.98. HRMS Calcd. for $C_{10}H_{13}NO^+$ $[M^+]$: 163.0997; found 163.1007.

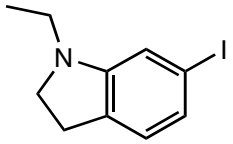


6-iodoindole (4-S2). According to the published protocol by Klapars and Buchwald[22], an oven dried, sealable glass tube was charged with a magnetic stirbar, 6-bromoindole (4.94 g, 25.2 mmol), freshly ground sodium iodide (7.55 g, 50.4 mmol), and copper(I) iodide (480 mg, 2.52 mmol). The vessel was then fitted with a rubber septum, evacuated under vacuum and backfilled with argon. This process was repeated three times. The vessel was then charged with 1,4-dioxane (25 mL) followed by N,N'-dimethylethylenediamine (0.58 mL, 5.40 mmol) via syringe. The rubber septum was removed and the reaction vessel immediately sealed tightly with a Teflon screw cap and heated to 110 ˚C for 22 hours. After cooling to room temperature, the reaction was

diluted with saturated aqueous NH_4Cl (30 mL) and extracted with dichloromethane (4 x 25 mL). The combined organic layers were washed with brine (30 mL) and dried over Na_2SO_4, then concentrated to a brown residue. The residue was triturated in hexanes and concentrated to provide **4-S2** (5.87 g, 99%) as a brown crystalline solid. This compound has been previously characterized[30]. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.18 (br s, 1H), 7.80 (s, 1H), 7.43 (s, 2H), 7.21 – 7.13 (t, *J* = 2.8 Hz, 2H), 6.56 (t, *J* = 2.2 Hz, 1H).

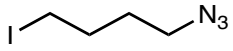


6-iodoindoline (4-2b). **4-S2** (5.86 g, 24.1 mmol) was dissolved in glacial acetic acid (100 mL) and cooled in an ice bath just until the solution began to become partially frozen. At this point, the solution was treated portion-wise with sodium cyanoborohydride (4.52 g, 72.3 mmol), then allowed to warm to room temperature and stir for 3 hours. The reaction mixture was concentrated to a thick residue, then diluted with 50 mL of saturated aqueous $NaHCO_3$. The resulting solution was extracted with ethyl acetate (3 x 40 mL). The combined organic layers were washed with saturated aqueous $NaHCO_3$ (30 mL) and brine (30 mL), then dried over anhydrous Na_2SO_4 and concentrated to a brown residue. Purification by flash chromatography (0-15% ethyl acetate/hexanes) afforded **4-2b** (4.78 g, 81%) as a white crystalline solid. This compound has been previously characterized[30]. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.10-6.95 (m, 2H), 6.87 (d, *J* = 7.5 Hz, 1H), 3.59 (td, *J* = 8.4, 2.9 Hz, 2H), 3.01 (t, *J* = 8.4 Hz, 2H).

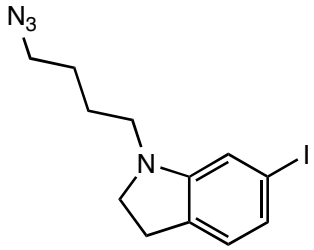


1-ethyl-6-iodoindoline (4-2c). **4-S2** (729 mg, 3.00 mmol) was dissolved in glacial acetic acid (15 mL) and treated with sodium borohydride (681 mg, 18.0 mmol) in small portions at room temperature. After stirring for 1 hour, the reaction was treated with sodium cyanoborohydride (377 mg, 6.00 mmol). The reaction was concentrated in vacuo to a thick residue, which was then diluted with 25 mL of saturated aqueous $NaHCO_3$. The resulting solution was extracted with ethyl acetate (25 mL, 3x). The combined organic layers were washed with saturated aqueous $NaHCO_3$ and brine (30 mL), then dried over anhydrous Na_2SO_4 and concentrated in vacuo to a

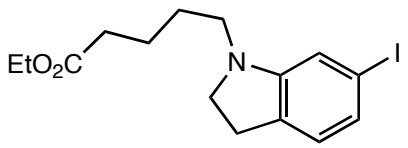
brown residue. Purification by flash chromatography (0-10% ethyl acetate/hexanes) afforded **4-2c** (590 mg, 76%) as a light yellow oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.97 (dd, *J* = 7.6, 1.1 Hz, 1H), 6.80 (d, *J* = 7.6 Hz, 1H), 6.78 (s, 1H), 3.38 (t, *J* = 8.2 Hz, 2H), 3.14 (q, *J* = 7.2 Hz, 2H), 2.93 (t, *J* = 8.2 Hz, 2H), 1.19 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 154.13, 130.53, 126.44, 126.27, 116.06, 92.76, 52.52, 42.98, 28.49, 12.15. HRMS (FAB) Calcd for $C_{10}H_{12}IN^+$ $[M]^+$: 273.0014; found 273.0011.



1-azido-4-iodobutane. Prepared according to the published protocol[31].

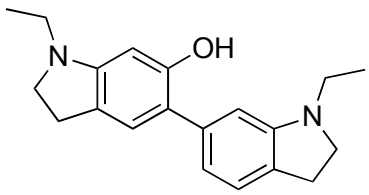


1-(4-azidobutyl)-6-iodoindoline (4-2d). 4-2b (294 mg, 1.20 mmol) was dissolved in DMF and treated with 1-azido-4-iodobutane (405 mg, 1.80 mmol) and sodium carbonate (382 mg, 3.60 mmol). The reaction was heated to 60 ˚C for 8 hours. After cooling to room temperature, the reaction was diluted with water (25 mL) and extracted with ethyl acetate (3 x 25 mL). The combined organic layers were washed with brine (30 mL) and dried over anhydrous Na_2SO_4, then concentrated to a light brown oil. Purification by flash chromatography (0-5% ethyl acetate/hexanes) provided 326 mg (80%) of **4-2d** as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.98 (dd, *J* = 7.6, 1.5 Hz, 1H), 6.81 (d, *J* = 7.6 Hz, 1H), 6.76 (d, *J* = 1.4 Hz, 1H), 3.49 – 3.29 (m, 4H), 3.08 (t, *J* = 6.8 Hz, 2H), 2.95 (t, *J* = 8.2 Hz, 2H), 1.77 – 1.65 (m, 4H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 154.39, 130.26, 126.58, 126.37, 115.77, 92.81, 53.39, 51.70, 48.71, 28.57, 27.01, 25.02. HRMS (FAB) Calcd. For $C_{12}H_{15}IN_4^+$ $[M^+]$: 342.0341; found 342.0328.

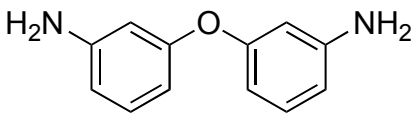


ethyl 5-(6-iodoindolin-1-yl)pentanoate (4-2e). **4-2b** (2.29 g, 9.34 mmol) was dissolved in DMF (15 mL) and treated with diisopropylethylamine (2.44 mL, 14.0 mmol), ethyl 5-bromovalerate (2.02 mL, 12.6 mmol) and potassium iodide (1.86 g, 11.2 mmol). The reaction was heated to 60

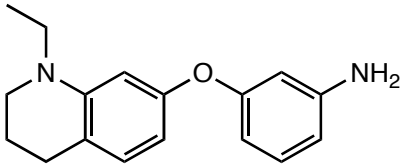
˚C for 12 hours. After cooling to room temperature, the reaction was diluted with water (30 mL) and concentrated aqueous NH_4Cl (10 mL), then extracted with ethyl acetate (3 x 30 mL). The combined organic layers were dried over anhydrous Na_2SO_4 and concentrated to a crude brown oil. Purification by flash chromatography (0-10% ethyl acetate/hexanes) provided **4-2e** (3.10 g, 89%) as a beige oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.95 (dd, *J* = 7.6, 1.5 Hz, 1H), 6.79 (d, *J* = 7.6 Hz, 1H), 6.74 (d, *J* = 1.3 Hz, 1H), 4.17 (q, *J* = 7.1 Hz, 2H), 3.38 (t, *J* = 8.3 Hz, 2H), 3.06 (t, *J* = 7.2 Hz, 2H), 2.93 (t, *J* = 8.3 Hz, 2H), 2.39 (t, *J* = 7.2 Hz, 2H), 1.79 – 1.70 (m, 2H), 1.69-1.60 (m, 2H), 1.30 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 173.79, 154.36, 130.26, 126.48, 126.27, 115.81, 92.74, 60.72, 53.32, 48.80, 34.42, 28.52, 27.08, 22.96, 14.68. HRMS (FAB) Calcd. For $C_{15}H_{20}INO_2^+$ [M^+]: 373.0539; found 373.0553.



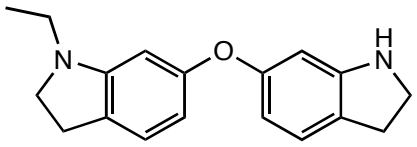
1,1'-diethyl-[5,6'-biindolin]-6-ol (4-S9). An oven dried sealable tube was purged with argon then charged with **4-1c** (163 mg, 1.00 mmol), palladium acetate (4 mg, 2.0 mol %), 2-(di-tert0butylphosphino)biphenyl (8 mg, 3.0 mol %), finely ground potassium phosphate (355 mg, 1.67 mmol), and a solution of **4-1c-triflate** (246 mg, 0.833 mmol) in anhydrous toluene (3 mL). The reaction was sealed and heated to 100 ˚C for 16 hours. After cooling, the reaction was filtered through celite, concentrated, and the major product **4-S9** purified by flash chromatography and isolated as a colorless oil (138 mg, 54%). The structure was confirmed by ^{1}H NMR and mass spectrometry. The ^{1}H NMR in $CDCl_3$ shows an asymmetrical product with the expected multiplicities for the aromatic protons (3 singlets, 2 doublets). An additional singlet was present at 6.50 ppm (phenol –OH). The corresponding singlet is not present in the spectrum when obtained in MeOD. Mass spectrometry (ESI) showed a correct mass of 309.5 (M+1). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.15 (d, *J* = 7.3 Hz, 1H), 6.97 (s, 1H), 6.69 (d, *J* = 7.1 Hz, 1H), 6.50 (s, 1H), 6.18 (s, 1H), 5.55 (s, 1H), 3.47-3.37 (m, 4H), 3.24 – 3.14 (m, 4H), 3.02 (t, *J* = 8.2 Hz, 2H), 2.95 (t, *J* = 8.1 Hz, 2H), 1.27-1.18 (m, 6H). ^{1}H NMR (400 MHz, MeOD) δ 7.05 (d, *J* = 7.5 Hz, 1H), 6.93 (s, 1H), 6.78 (d, *J* = 7.5, 1H), 6.71 (s, 1H), 6.13 (s, 1H), 3.30 (td, *J* = 8.1, 1.8 Hz, 4H), 3.12 (dq, *J* = 12.0, 7.2 Hz, 4H), 2.92 (t, *J* = 8.1 Hz, 2H), 2.86 (t, *J* = 8.0 Hz, 2H), 1.20 (t, *J* = 7.2 Hz, 6H).



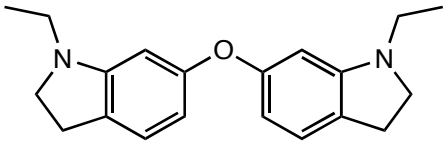
3,3'-oxydianiline (4-3a). Following the general procedure, 3-aminophenol (**4-1a,** 1.32 g, 12.0 mmol) and 3-iodoaniline (**4-2a**, 2.18 g, 10.0 mmol) were coupled to provide **4-3a** (1.80 g, 90%) as a dense tan crystalline solid. This compound has been previously characterized[21]. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.08 (t, *J* = 8.0 Hz, 2H), 6.42 (m, 4H), 6.34 (s, 2H), 3.67 (br s, 4H).



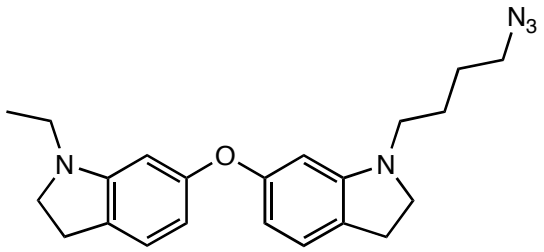
3-(1-ethyl-1,2,3,4-tetrahydroquinolin-7-yloxy)aniline (4-3b). Following the general procedure, phenol **4-1b** (1.20 g, 12.0 mmol) and 3-iodoaniline (**4-2a,** 1.23 g, 10.0 mmol) were coupled to provide **4-3b** (1.24 g, 82%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.13 (t, *J* = 8.0 Hz, 1H), 6.94 (d, *J* = 8.0 Hz, 1H), 6.49 (ddd, *J* = 8.2, 2.2, 0.9 Hz, 1H), 6.44-6.38 (m, 3H), 6.29 (dd, *J* = 8.0, 2.2 Hz, 1H), 3.70 (s, 2H), 3.39-3.29 (m, 4H), 2.80 (t, *J* = 6.4 Hz, 2H), 2.03 (dq, *J* = 9.0, 6.2 Hz, 2H), 1.18 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.72, 156.50, 148.38, 146.53, 130.54, 130.12, 118.25, 109.77, 108.63, 106.63, 105.21, 102.69, 48.64, 45.84, 28.05, 22.80, 11.18. HRMS (FAB+) Calcd. For $C_{17}H_{20}N_2O^+$ [M^+]: 268.1576; found 268.1583.



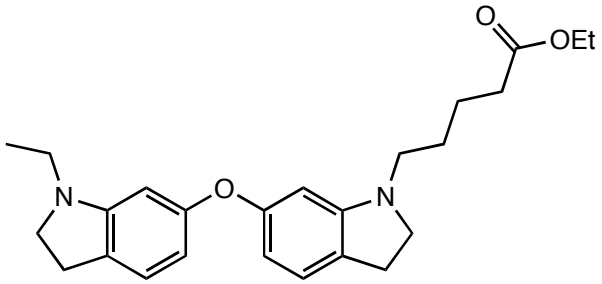
1-ethyl-6-(indolin-6-yloxy)indoline (4-3c). Following the general procedure, phenol **4-1c** (196 mg, 1.20 mmol) and aryl iodide **4-2b** (245 mg, 1.00 mmol) were coupled to provide **4-3c** (251 mg, 90%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.05 (d, *J* = 7.9 Hz, 1H), 7.00 (d, *J* = 7.9 Hz, 1H), 6.40 (dd, *J* = 7.9, 2.2 Hz, 1H), 6.36 (d, *J* = 2.1 Hz, 1H), 6.31 (dd, *J* = 7.8, 2.2 Hz, 1H), 6.23 (d, *J* = 2.1 Hz, 1H), 4.32 (br s, 1H), 3.61 (t, *J* = 8.3 Hz, 2H), 3.41 (t, *J* = 8.3 Hz, 2H), 3.12 (q, *J* = 7.2 Hz, 2H), 3.03 (t, *J* = 8.2 Hz, 2H), 2.96 (t, *J* = 8.0 Hz, 2H), 1.20 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 158.29, 157.99, 154.12, 153.15, 125.23, 124.91, 109.08, 108.04, 104.17, 101.00, 99.64, 96.18, 53.19, 48.36, 43.37, 29.53, 28.32, 12.20. HRMS (FAB) Calcd. For $C_{18}H_{20}N_2O^+$ [M^+]: 280.1576; found 280.1579.



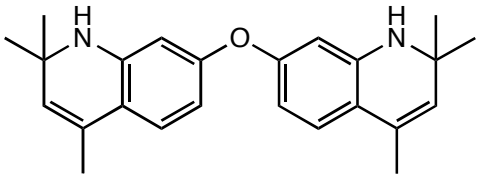
6,6'-oxybis(1-ethylindoline) (4-3d). Following the general procedure, phenol **4-1c** (412 mg, 2.53 mmol) and aryl iodide **4-2c** (575 mg, 2.10 mmol) were coupled to provide **4-3d** (520 mg, 80%) as a clear colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.98 (d, *J* = 7.8 Hz, 1H), 6.29 (d, *J* = 7.8 Hz, 1H), 6.22 (s, 1H), 3.40 (t, *J* = 8.2 Hz, 2H), 3.11 (q, *J* = 7.2 Hz, 2H), 2.95 (t, *J* = 8.2 Hz, 2), 1.18 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 158.26, 154.02, 125.10, 124.85, 107.64, 99.27, 53.17, 43.33, 28.31, 12.19. HRMS (FAB) Calcd. For $C_{20}H_{24}N_2O^+$ [M^+]: 308.1889; found 308.1897.



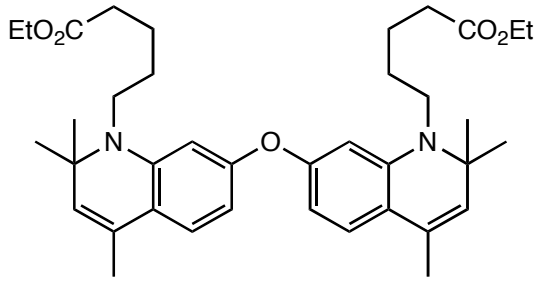
1-(4-azidobutyl)-6-(1-ethylindolin-6-yloxy)indoline (4-3e). Following the general procedure, the phenol **4-1c** (166 mg, 1.02 mmol) and the aryl iodide **4-2d** (290 mg, 0.848 mmol) were coupled to provide **4-3e** (245 mg, 77%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.01 (d, *J* = 3.5 Hz, 1H), 6.99 (d, *J* = 3.5 Hz, 1H), 6.32 (d, *J* = 2.3 Hz, 1H), 6.30 (t, *J* = 2.4 Hz, 1H), 6.24 (d, *J* = 2.1 Hz, 1H), 6.22 (d, *J* = 2.1 Hz, 1H), 3.43 (t, *J* = 8.2 Hz, 4H), 3.39 – 3.33 (m, 2H), 3.14 (q, *J* = 7.2 Hz, 2H), 3.10-3.05 (m, 2H), 2.98 (td, *J* = 8.1, 2.5 Hz, 4H), 1.78 – 1.67 (m, 4H), 1.21 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 158.36, 158.18, 154.32, 154.05, 125.21, 124.92, 124.87, 124.72, 107.67, 107.54, 99.32, 98.86, 54.00, 53.17, 51.74, 48.98, 43.33, 28.36, 28.32, 27.02, 25.04, 12.19. HRMS (FAB) Calcd. For $C_{22}H_{27}N_5O^+$ [M^+]: 377.2216; found 377.2223.



ethyl 5-(6-(1-ethylindolin-6-yloxy)indolin-1-yl)pentanoate (4-3f). Following the general procedure, the phenol **4-1c** (1.60 g, 9.77 mmol) and the aryl iodide **4-2e** (3.04, 8.14 mmol) were coupled to provide **4-3f** (2.67 g, 80%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.98 (d, *J* = 7.8 Hz, 1H), 6.29 (d, *J* = 7.8 Hz, 1H), 6.22 (s, 1H), 3.40 (t, *J* = 8.2 Hz, 2H), 3.11 (q, *J* = 7.2 Hz, 2H), 2.95 (t, *J* = 8.2 Hz, 2), 1.18 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.85, 158.28, 158.20, 154.34, 125.14, 124.83, 124.77, 107.69, 107.44, 99.33, 98.93, 60.66, 53.94, 53.15, 49.08, 43.35, 34.48, 28.32, 28.29, 27.16, 23.01, 14.65, 12.15. HRMS (FAB) Calcd. For $C_{25}H_{32}N_2O_3^+$ [M^+]: 408.2413; found 408.2408.

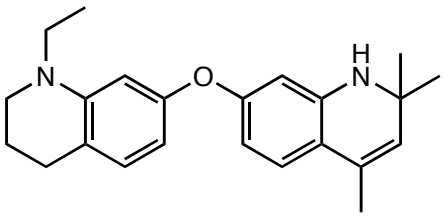


7,7'-oxybis(2,2,4-trimethyl-1,2-dihydroquinoline) (4-S3). **4-3a** (400 mg, 2.00 mmol) was dissolved in 2,2-dimethoxypropane (5 mL) and treated with p-toluenesulfonic acid monohydrate (190 mg, 1.00 mmol) and stirred at room temperature for 5 days. The reaction was quenched by the addition of 4% aq. $NaHCO_3$ (30 mL) and extracted with ethyl acetate (3 x 25 mL). The combined organic layers were washed with brine, dried over anhydrous Na_2SO_4 and concentrated to a crude residue. Purified by flash chromatography to provide 340 mg of **4-S3** as a white crystalline solid (47%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.03 (d, *J* = 8.2 Hz, 1H), 6.35 (dd, *J* = 8.2, 1.6 Hz, 1H), 6.13 (d, *J* = 1.6 Hz, 1H), 5.26 (s, 1H), 3.63 (br s, 1H), 2.01 (s, 3H), 1.28 (s, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 158.10, 145.01, 128.38, 127.32, 125.10, 117.36, 108.00, 103.55, 52.39, 31.59, 19.07. HRMS (FAB+) Calcd. For $C_{24}H_{29}N_2O^+$ [M+1]: 361.2274; found 361.2280.

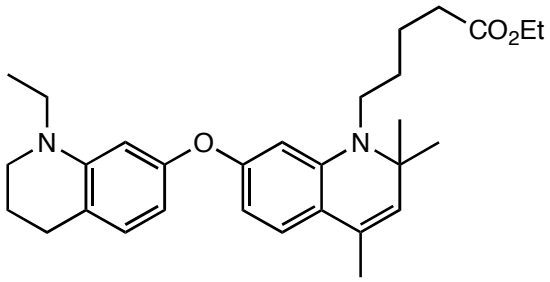


diethyl 5,5'-(7,7'-oxybis(2,2,4-trimethylquinoline-7,1(2*H*)-diyl))dipentanoate (4-S4). A solution of **4-S3** (270 mg, 0.75 mmol) in anhydrous acetonitrile (4 mL) was treated with sodium iodide (112 mg, 0.75 mmol), sodium carbonate (476 mg, 4.50 mmol), and ethyl 5-bromovalerate (0.60 mL, 3.75 mmol), then heated to reflux. After 24 hours, the reaction was charged with

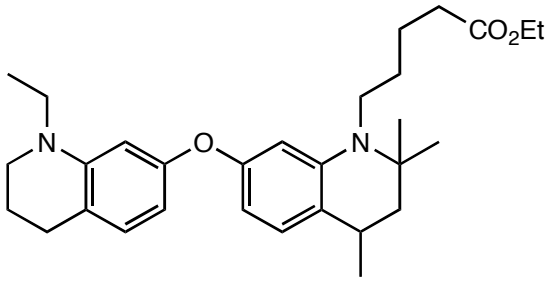
additional ethyl 2-bromovalerate (0.30 mL, 1.9 mmol) and sodium carbonate (238 mg, 2.25 mmol). After an additional 24 hours of heating, TLC of the reaction indicated a ~3:1 mixture of the desired bis-alkylated product to the mono-alkylated product. The reaction was then quenched by the addition of water (30 mL) and extracted with ethyl acetate (3 x 25 mL). The combined organic layers were washed with brine and dried over anhydrous Na_2SO_4 and concentrated to a crude residue. Purified by flash chromatography to provide 230 mg of **4-S4** as a colorless oil (50%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.98 (d, *J* = 8.3 Hz, 2H), 6.25 (dd, *J* = 8.3, 2.1 Hz, 2H), 6.15 (d, *J* = 2.2 Hz, 2H), 5.17 (d, *J* = 1.2 Hz, 2H), 4.14 (q, *J* = 7.1 Hz, 4H), 3.17 (t, *J* = 7.0 Hz, 4H), 2.32 (t, *J* = 6.9 Hz, 4H), 1.97 (d, *J* = 1.1 Hz, 6H), 1.69 – 1.57 (m, 8H), 1.32 (s, 12H), 1.27 (t, *J* = 7.1 Hz, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 173.75, 158.44, 145.65, 128.10, 127.83, 124.81, 118.69, 105.63, 102.00, 60.65, 57.26, 44.25, 34.46, 28.88, 28.29, 22.96, 19.13, 14.64. HRMS (FAB+) Calcd. For $C_{38}H_{53}N_2O_5^+$ [M+1]: 617.3949; found 617.3947.



7-(1-ethyl-1,2,3,4-tetrahydroquinolin-7-yloxy)-2,2,4-trimethyl-1,2-dihydroquinoline (4-S5). A solution of **4-3b** (1.20 g, 4.47 mmol) in acetone (80 mL) was treated with p-toluenesulfonic acid monohydrate (213 mg, 1.12 mmol) and stirred at room temperature for 48 hours. The reaction was then concentrated to a residue under reduced pressure, brough up 4% aq. $NaHCO_3$ (30 mL) and extracted with ethyl acetate (3 x 25 mL). The combined organics were washed with brine and dried over anhydrous Na_2SO_4 and concentrated to a crude residue. Purified by flash chromatography to provide 1.17 g of **4-S5** as a white foamy solid (75%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.03 (d, *J* = 8.4 Hz, 1H), 6.91 (d, *J* = 8.0 Hz, 1H), 6.38 (d, *J* = 2.3 Hz, 1H), 6.35 (dd, *J* = 8.4, 2.4 Hz, 1H), 6.27 (dd, *J* = 8.0, 2.3 Hz, 1H), 6.14 (d, *J* = 2.4 Hz, 1H), 3.71 (s, 1H), 3.37-3.27 (m, 4H), 2.77 (t, *J* = 6.4 Hz, 2H), 2.06 – 1.99 (m, 5H), 1.31 (s, 6H), 1.16 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.07, 156.45, 146.44, 145.01, 130.03, 128.60, 126.92, 125.01, 118.14, 116.89, 107.08, 106.71, 102.71, 102.65, 52.36, 48.61, 45.81, 31.52, 28.02, 22.78, 19.07, 11.15. HRMS (FAB+) Calcd. For $C_{23}H_{28}N_2O^+$ [M^+]: 348.2202; found 348.2186.

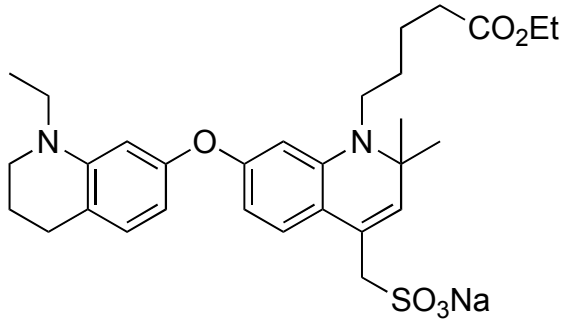


ethyl 5-(7-(1-ethyl-1,2,3,4-tetrahydroquinolin-7-yloxy)-2,2,4-trimethylquinolin-1(2*H*)-yl)pentanoate (4-S6). According to the protocol used to synthesize **4-S4**, **4-S5** (1.05 g, 3.00 mmol) was subjected to alkylation and purified by flash chromatography to yield 845 mg of **4-S6** as a colorless oil (59%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.97 (d, *J* = 8.3 Hz, 1H), 6.88 (d, *J* = 8.0 Hz, 1H), 6.36 (d, *J* = 2.3 Hz, 1H), 6.25 (dd, *J* = 4.3, 2.3 Hz, 1H), 6.23 (dd, *J* = 4.6, 2.3 Hz, 1H), 6.11 (d, *J* = 2.2 Hz, 1H), 5.15 (d, *J* = 1.3 Hz, 1H), 4.16 (q, *J* = 7.1 Hz, 2H), 3.35 – 3.25 (m, 4H), 3.15 (t, *J* = 7.2 Hz, 2H), 2.74 (t, *J* = 6.3 Hz, 2H), 2.31 (t, *J* = 7.1 Hz, 2H), 2.03 – 1.93 (m, 2H), 1.97 (s, 3H) 1.67-1.55 (m, 4H), 1.32 (s, 6H), 1.29 (t, *J* = 7.1 Hz, 3H), 1.13 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 173.80, 164.92, 159.31, 156.47, 146.37, 145.54, 129.94, 127.84, 124.81, 118.14, 118.02, 106.52, 104.81, 102.60, 101.17, 60.66, 57.25, 48.58, 45.78, 44.24, 34.47, 28.90, 28.28, 27.95, 22.94, 22.75, 19.13, 14.65, 11.13. HRMS (FAB+) Calcd. For $C_{30}H_{41}N_2O_3^+$ [M+1]: 477.3112; found 477.3129.

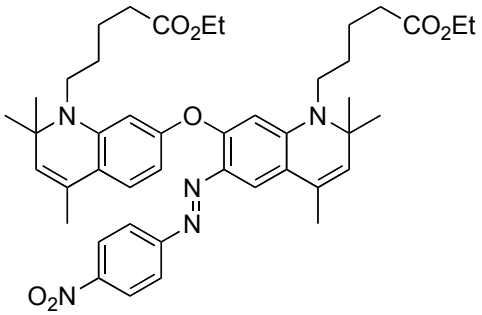


ethyl-5-(7-(1-ethyl-1,2,3,4-tetrahydroquinolin-7-yloxy)-2,2,4-trimethyl-3,4-dihydroquinolin-1(2*H*)-yl)pentanoate (4-S7). A round bottom flask charged with a magnetic stirbar, **4-S6** (185 mg, 0.39 mmol) and ethyl acetate (6 mL) was evacuated under vacuum and backfilled with argon gas. This process was repeated three times. The vessel was then charged with 10% palladium on carbon (41 mg, 0.039 mmol), evacuated with vacuum, and backfilled with hydrogen gas. This process was repeated five times, then the reaction left to stir overnight under an atmosphere of hydrogen gas (balloon). The next morning, the reaction was filtered through a pad of celite and concentrated under reduced pressure to provide 180 mg of **4-S7** as a colorless oil (97%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.08 (d, *J* = 8.3 Hz, 1H), 6.91 (d, *J* = 8.0 Hz, 1H), 6.41 (s, 1H), 6.31 (d, *J* =

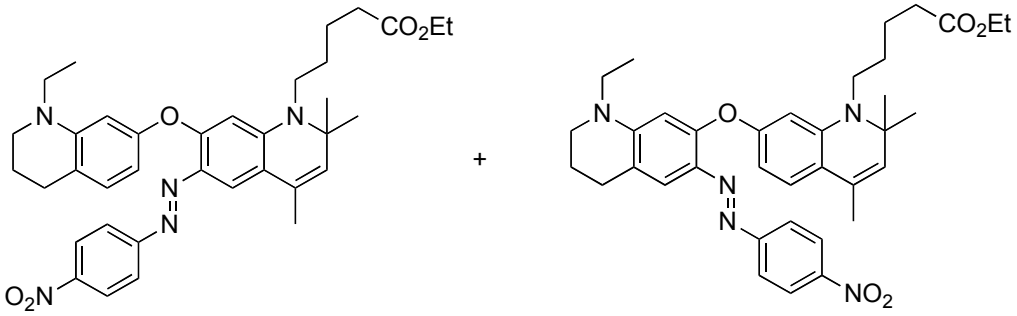
8.3 Hz, 1H), 6.28 (d, *J* = 8.0, 1H), 6.21 (s, 1H), 4.20 (q, *J* = 7.1 Hz, 2H), 3.40 – 3.25 (m, 5H), 3.08 - 2.97 (m, 1H), 2.95 – 2.84 (m, 1H), 2.77 (t, *J* = 6.2 Hz, 2H), 2.35 (m, 2H), 2.05 – 1.97 (m, 2H), 1.77 (dd, *J* = 12.9, 4.5 Hz, 1H), 1.70 - 1.54 (m, 5H), 1.43 – 1.28 (m, 9H), 1.23 (s, 3H), 1.17 (t, *J* = 7.0 Hz, 3H ^{13}C NMR (300 MHz, $CDCl_3$) δ 173.91, 157.88, 156.64, 146.37, 146.30, 129.98, 126.73, 122.43, 117.89, 106.58, 104.62, 102.75, 101.60, 60.72, 54.84, 48.62, 47.52, 45.82, 45.31, 34.54, 30.08, 29.01, 28.00, 27.28, 25.50, 22.94, 22.81, 20.56, 14.74, 11.16. HRMS (FAB+) Calcd. For $C_{30}H_{42}N_2O_3^+$ [M^+]: 478.3195; found 478.3210.



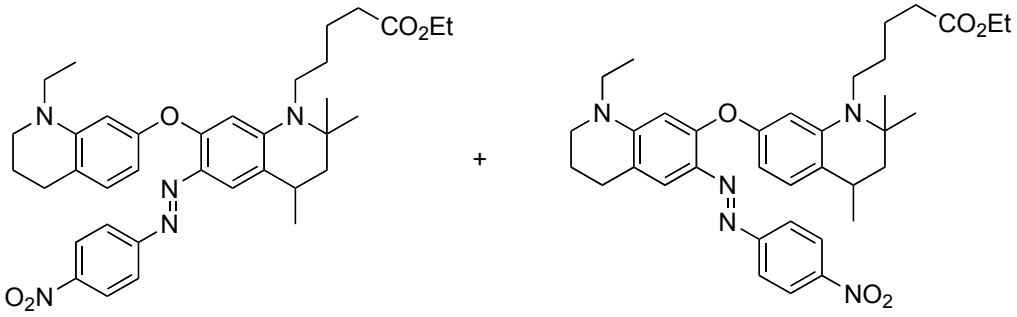
(1-(5-ethoxy-5-oxopentyl)-7-(1-ethyl-1,2,3,4-tetrahydroquinolin-7-yloxy)-2,2-dimethyl-1,2-dihydroquinolin-4-yl)methanesulfonate (4-S8) sodium salt. To a solution of fuming sulfuric acid (20% SO_3, 1.2 mL) in sulfuric acid (4 mL) cooled in a brine ice bath was added **4-S6** (954 mg, 2.00 mmol). The mixture was stirred on ice for 30 minutes then allowed to warm to room temperature and stir for 12 hours at room temperature. The reaction was then poured into a beaker of ice (50 grams) and the resulting solution, cooled in an ice bath, was neutralized to pH 7 with 10% sodium hydroxide. The precipitated sodium sulfate solids were filtered, and the resulting aqueous filtrate was concentrated under reduced pressure to white solids. This material was boilded gently in 30 mL of ethanol, filtered, and the solids washed with additional hot ethanol. The filtrate was concentrated under reduced pressure and purified by flash chromatography to provide 825 mg of **4-S8** as a light pink oil (72%). ^{1}H NMR (400 MHz, MeOD) δ 7.28 (d, *J* = 8.4 Hz, 1H), 6.83 (d, *J* = 8.0 Hz, 1H), 6.27 (d, *J* = 1.8 Hz, 1H), 6.18 (ddd, *J* = 14.3, 8.2, 1.9 Hz, 2H), 5.98 (d, *J* = 1.7 Hz, 1H), 5.50 (s, 1H), 4.11 (t, *J* = 7.2 Hz, 2H), 3.88 (s, 2H), 3.25 (m, 4H), 3.10 (m, 2H), 2.68 (t, *J* = 6.2 Hz, 2H), 2.24 (t, *J* = 6.6 Hz, 2H), 1.96 – 1.87 (m, 2H), 1.78 – 1.64 (m, 2H), 1.64 – 1.56 (m, 2H), 1.33 (s, 6H), 1.25 (t, *J* = 7.1 Hz, 3H), 1.08 (t, *J* = 7.0 Hz, 3H). ^{13}C NMR (300 MHz, MeOD) δ 176.67, 174.15, 159.61, 156.29, 146.28, 145.58, 132.49, 129.71, 125.84, 124.63, 118.30, 116.59, 106.86, 104.70, 102.72, 100.70, 67.60, 61.55, 60.52, 57.03, 53.82, 48.28, 45.34, 43.91, 33.89, 33.62, 32.04, 28.77, 27.89, 27.67, 27.37, 22.56, 22.46, 21.54, 13.74, 10.10. HRMS (FAB+) Calcd. For $C_{30}H_{39}N_2O_6SNa_2^+$ [M+2Na]: 601.2324; found 601.2307.



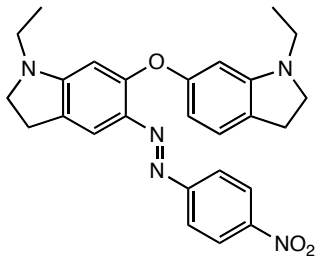
Diazene 4-4a. Following the general procedure, **4-S4** (205 mg, 0.332 mmol) was reacted to give 167 mg of **4-4a** as a deep purple residue (66%).



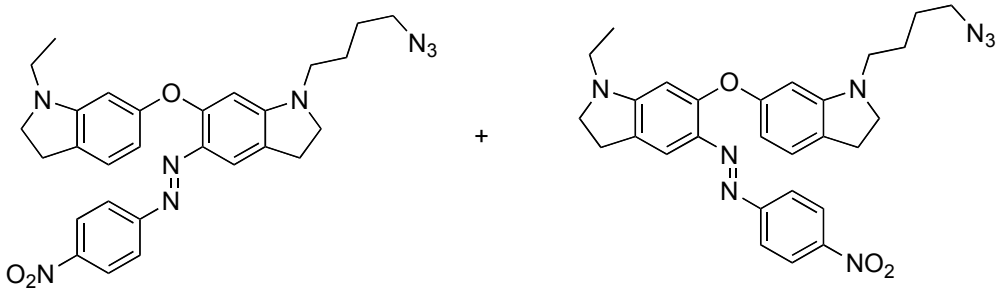
Diazenes 4-4b and 4-5b. Following the general procedure, **4-S6** (185 mg, 0.388 mmol) was reacted to give 155 mg of a mixture of **4-4b** and **4-5b** as a deep purple residue (69%).



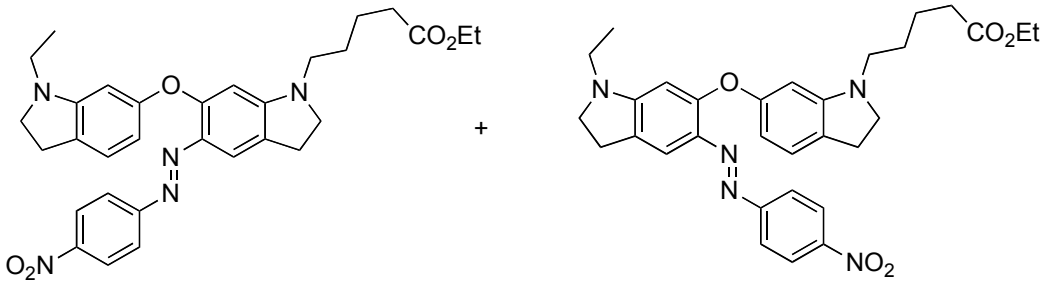
Diazenes 4-4c and 4-5c. Following the general procedure, **4-S7** (180 mg, 0.376 mmol) was reacted to give 170 mg of a mixture of **4-4c** and **4-5c** as a deep purple residue (78%).



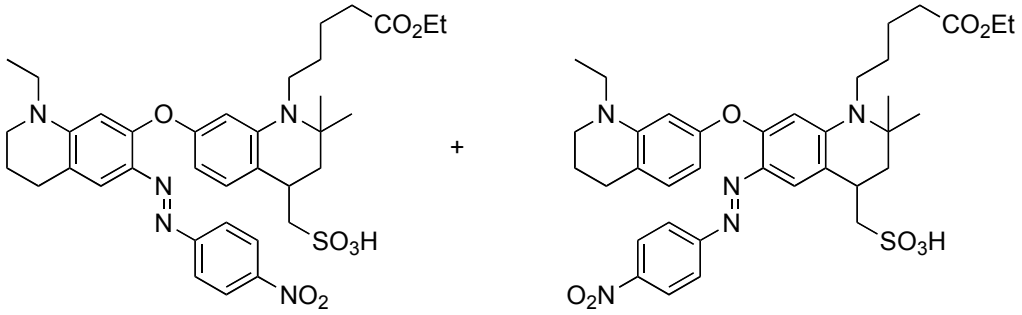
Diazene 4-4d. Following the general procedure, **4-3d** (230 mg, 0.756 mmol) was reacted to give 270 mg of **4-4d** as a deep red/bronze solid (79%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.24 (d, *J* = 9.1 Hz, 2H), 7.79-7.71 (m, 3H), 7.00 (d, *J* = 7.6 Hz, 1H), 6.34 (d, *J* = 7.9 Hz, 1H), 6.30 (s, 1H), 6.06 (s, 1H), 3.67 (t, *J* = 8.2 Hz, 2H), 3.42 (t, *J* = 8.2 Hz, 2H), 3.27 (t, *J* = 7.1 Hz, 2H), 3.13 (t, *J* = 7.2 Hz, 2H), 3.08 (t, *J* = 8.1 Hz, 2H), 2.97 (t, *J* = 8.2 Hz, 2H), 1.26-1.14 (m, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 160.66, 158.73, 158.04, 157.61, 154.21, 146.99, 135.85, 127.05, 125.56, 124.98, 122.91, 113.32, 107.55, 99.21, 96.07, 53.18, 51.95, 43.28, 41.59, 28.31, 26.97, 12.11. HRMS (FAB+) Calcd. For $C_{26}H_{27}N_5O_3^+$ [M^+]: 457.2114; found 457.2131.



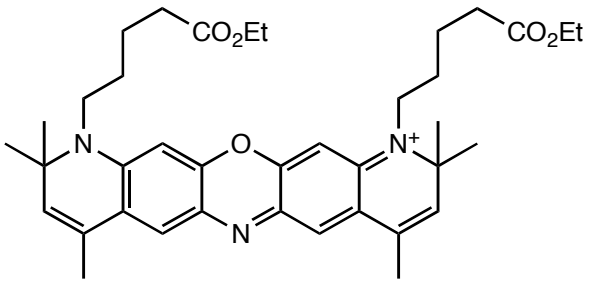
Diazenes 4-4e and 4-5e. Following the general procedure, **4-3e** (240 mg, 0.636 mmol) was reacted to give 237 mg of a mixture of **4-4e** and **4-5e** as a deep purple film (71%).



Diazenes 4-4f and 4-5f. Following the general procedure, **4-3f** (1.42 g, 3.47 mmol) was reacted to give 1.46 g of a mixture of **4-4f** and **4-5f** as a deep purple residue (76%).

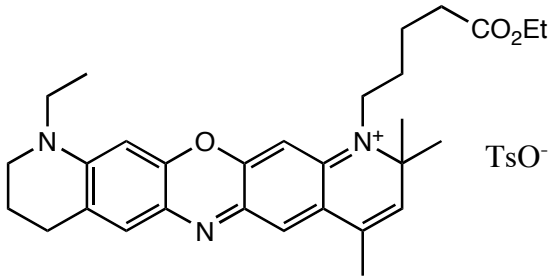


Diazenes 4-4g and 4-5g. A round bottom flask charged with a magnetic stir-bar, **4-S8** (150 mg, 0.26 mmol) and methanol (10 mL) was evacuated under vacuum and backfilled with argon gas. This process was repeated three times. The vessel was then charged with 10% palladium on carbon (30 mg, 0.026 mmol), evacuated with vacuum, and backfilled with hydrogen gas. This process was repeated five times, then the reaction left to stir overnight under an atmosphere of hydrogen gas (balloon). The next morning, the reaction was filtered through a pad of celite and concentrated under reduced pressure to tan oil. The resulting oil was then brought up in methanol (2 mL) and cooled in an ice bath for 10 minutes. The solution was then treated with 2N aqueous HCl (5 mL, pre-cooled in ice bath) followed by the slow addition of p-nitrobenzenediazonium tetrafluoroborate (62 mg, 0.26 mmol). After stirring on ice for 1.5 hours, the reaction was diluted with water (20 mL) and extracted directly without neutralization (DCM, 3 x 25 mL). The combined organic layers were dried over Na_2SO_4 and concentrated in vacuo to a deep red-purple residue. Flash chromatography (5-30% dichloromethane/methanol) provided the regioisomeric mixture of the diazenes **4-4g** and **4-5g** (115 mg, 63%).

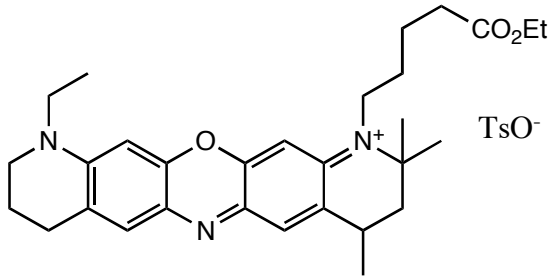


Oxazine 4-6a. Following the general procedure, **4-4a** (91 mg, 0.119 mmol) was cyclized to provide 70 mg of **4-6a** as a purple film (74%). ^{1}H NMR (400 MHz, MeOD) δ 7.70 (d, J = 8.1 Hz, 2H), 7.42 (s, 2H), 7.19 (d, J = 8.0 Hz, 2H), 6.74 (s, 2H), 5.85 (s, 2H), 4.18 (q, J = 7.1 Hz, 4H), 3.73 (br s, 4H), 2.48 (br s, 4H), 2.35 (s, 3H), 2.13 (s, 6H), 1.83 (br s, 8H), 1.59 (s, 12H), 1.28 (t, J

= 7.1 Hz, 6H). ^{13}C NMR (300 MHz, MeOD) δ 173.92, 153.45, 149.98, 142.77, 140.44, 135.13, 134.47, 128.71, 126.69, 125.96, 125.78, 124.72, 96.03, 61.76, 60.58, 33.42, 28.35, 27.41, 22.16, 20.36, 17.64, 13.66. HRMS (FAB) Calcd. for $C_{38}H_{50}N_3O_5^+$ [M^+]: 628.3745; found 628.3775.

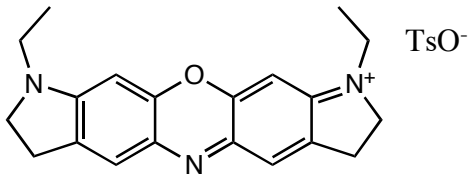


Oxazine 4-6b. Following the general procedure, the mixture of regioisomeric diazenes **4-4b** and **4-5b** (140 mg, 0.241 mmol) was cyclized to provide 135 mg of **4-6b** as a deep blue film (85%). ^{1}H NMR (400 MHz, MeOD) δ 7.70 (d, *J* = 8.0 Hz, 2H), 7.43 (s, 1H), 7.40 (s, 1H), 7.19 (d, *J* = 8.0 Hz, 2H), 6.90 (s, 1H), 6.68 (s, 1H), 5.82 (s, 1H), 4.18 (q, *J* = 7.1 Hz, 2H), 3.80 – 3.64 (m, 6H), 2.94 (t, *J* = 5.9 Hz, 2H), 2.48 (br s, 2H), 2.35 (s, 3H), 2.11 (s, 3H), 2.09 – 2.02 (m, 2H), 1.82 (br s, 4H), 1.58 (s, 6H), 1.37 (t, *J* = 7.0 Hz, 3H), 1.28 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, MeOD) δ 173.95, 154.47, 153.30, 149.94, 148.86, 142.75, 140.48, 135.05, 134.19, 130.69, 129.82, 128.72, 126.26, 125.96, 125.76, 124.74, 95.80, 95.31, 61.59, 60.58, 50.26, 46.16, 33.39, 28.40, 27.41, 27.35, 22.14, 20.93, 20.35, 17.64, 13.65, 10.79. HRMS (FAB) Calcd. For $C_{30}H_{38}N_3O_3^+$ [M^+]: 488.2908; found 488.2897.

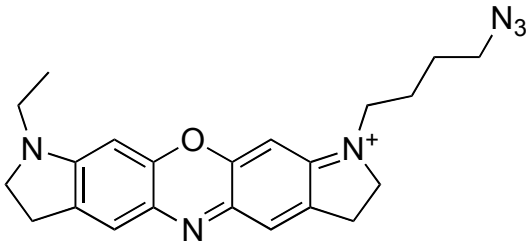


Oxazine 4-6c. Following the general procedure, the mixture of regioisomeric diazenes **4-4c** and **4-5c** (150 mg, 0.257 mmol) was cyclized to provide 148 mg of **4-6d** as a deep blue film (87%). ^{1}H NMR (400 MHz, MeOD) δ 7.68 (d, *J* = 8.1 Hz, 2H), 7.50 (s, 1H), 7.41 (s, 1H), 7.15 (d, *J* = 8.0 Hz, 2H), 6.88 (s, 1H), 6.66 (s, 1H), 4.17 (q, *J* = 7.1 Hz, 2H), 3.70 (t, *J* = 5.6 Hz, 4H), 3.57 – 3.44 (m, 1H), 2.98 (d, *J* = 6.5 Hz, 1H), 2.95 – 2.90 (m, 2H), 2.47 (br s, 2H), 2.31 (s, 3H), 2.10 – 1.97 (m, 3H), 1.79 (br s, 4H), 1.53 (s, 3H), 1.45 (s, 6H), 1.36 (t, *J* = 6.9 Hz, 3H), 1.28 (t, *J* = 7.1 Hz, 4H). ^{13}C NMR (300 MHz, MeOD) δ 173.90, 154.75, 154.32, 149.04, 148.01, 142.96, 140.32,

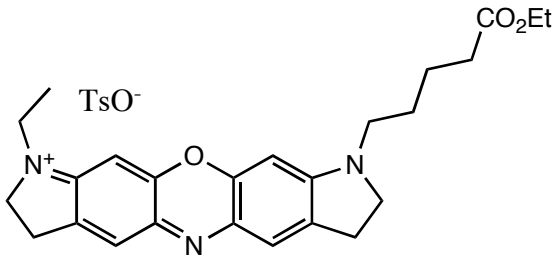
135.02, 134.20, 133.45, 130.84, 129.89, 128.68, 128.13, 125.98, 96.56, 95.30, 60.56, 59.25, 50.34, 46.62, 44.47, 33.44, 28.40, 27.75, 27.40, 27.16, 25.26, 22.27, 20.90, 20.44, 18.49, 13.73, 10.90. HRMS (FAB) Calcd. For $C_{32}H_{38}N_3O_3^+$ [M^+]: 490.3064; found 490.3079.



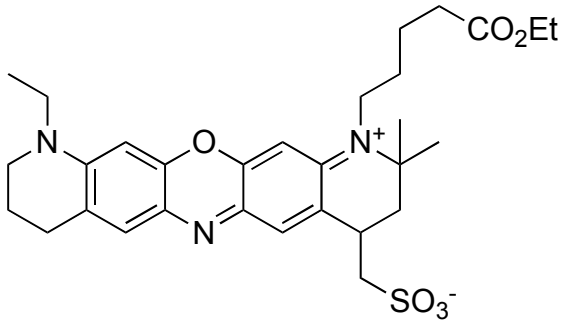
Oxazine 4-6d. Following the general procedure, diazene **4-4d** (108 mg, 0.236 mmol) was cyclized to provide 105 mg of **4-6d** as a deep blue film (90%). ^{1}H NMR (400 MHz, MeOD) δ 7.71 (d, *J* = 8.2 Hz, 2H), 7.40 (s, 2H), 7.22 (d, *J* = 8.0 Hz, 2H), 6.68 (s, 2H), 4.03 (t, *J* = 7.2 Hz, 4H), 3.66 (q, *J* = 7.3 Hz, 4H), 3.29 (t *J* = 7.2 Hz, 4H), 2.36 (s, 3H), 1.35 (t, *J* = 7.3 Hz, 6H). ^{13}C NMR (300 MHz, MeOD) δ 159.56, 150.80, 142.51, 140.64, 137.96, 134.66, 128.75, 125.96, 90.43, 52.73, 41.92, 26.09, 20.29, 11.16. HRMS (FAB) Calcd. for $C_{20}H_{22}N_3O^+$ [M^+]: 320.1757; found 320.1764.



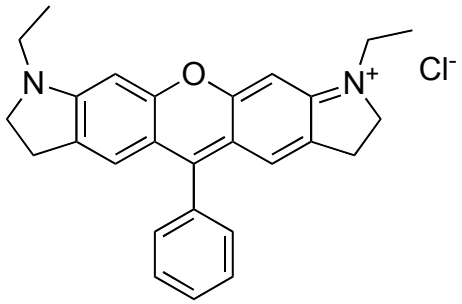
Oxazine 4-6e. Following the general procedure, the mixture of regioisomeric diazenes **4-4e** and **4-5e** (157 mg, 0.298 mmol) was cyclized to provide 130 mg of **4-6e** as a dark blue film (78%). ^{1}H NMR (400 MHz, MeOD) δ 7.51 – 7.41 (m, 2H), 6.74 (s, 1H), 6.72 (s, 1H), 4.05 (q, *J* = 6.6 Hz, 4H), 3.75-3.62 (m, 4H), 3.43 (t, *J* = 6.6 Hz, 2H), 3.31 (m, 4H), 1.92 – 1.80 (m, 2H), 1.77-1.66 (m, 2H), 1.36 (t, *J* = 7.3 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.96, 159.73, 150.99, 150.79, 138.22, 137.65, 135.03, 134.58, 126.07, 90.55, 90.43, 53.25, 52.78, 51.08, 46.74, 41.94, 26.26, 26.14, 26.07, 24.40, 11.15. HRMS (FAB) Calcd. for $C_{22}H_{25}N_6O^+$ [M^+]: 389.2084; found 389.2090.



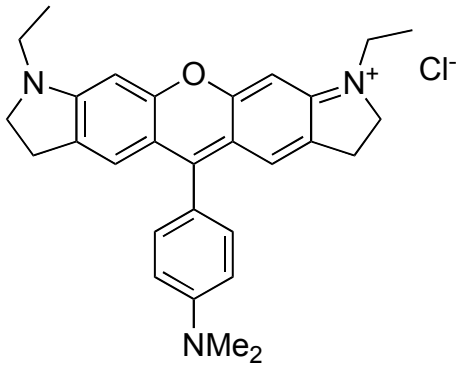
Oxazine 4-6f. Following the general procedure, the mixture of regioisomeric diazenes **4-4d** and **4-5d** (1.05 g, 1.88 mmol) were cyclized to provide 1.04 g of **4-7f** as a metallic red amorphous solid (94%). ^{1}H NMR (400 MHz, MeOD) δ 7.69 (d, *J* = 8.0 Hz, 2H), 7.24 (s, 2H), 7.16 (d, *J* = 8.0 Hz, 2H), 6.60 (s, 1H), 6.56 (s, 1H), 4.14 (q, *J* = 7.1 Hz, 2H), 3.97 (t, *J* = 6.3 Hz, 4H), 3.63 – 3.51 (m, 4H), 3.21 (br s, 4H), 2.42 (t, *J* = 6.9 Hz, 2H), 2.31 (s, 3H), 1.81 – 1.65 (m, 4H), 1.31 (t, *J* = 7.2 Hz, 3H), 1.26 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, MeOD) δ 173.95, 159.77, 159.43, 150.48, 150.37, 143.03, 140.38, 138.04, 137.62, 134.61, 134.32, 128.73, 125.98, 90.66, 90.52, 60.54, 53.29, 52.80, 47.00, 42.04, 33.48, 26.50, 26.18, 26.14, 22.28, 20.41, 13.67, 11.35. HRMS (FAB) Calcd. for $C_{25}H_{30}N_3O_3^+$ [M^+]: 420.2282; found 420.2291.



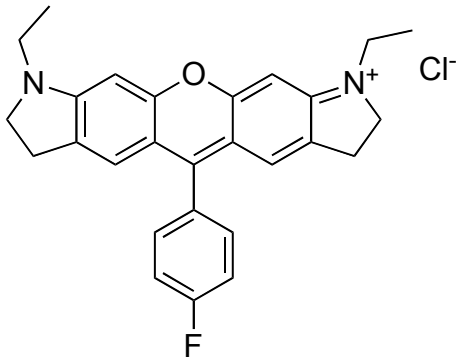
Oxazine 4-6g. Following the general procedure, the mixture of regioisomeric diazenes **4-4g** and **4-5g** (102 mg, 0.144 mmol) were cyclized to provide 68 mg of **4-7g** as a deep blue residue (83%). ^{1}H NMR (400 MHz, MeOD) δ 7.69 (s, 1H), 7.41 (s, 1H), 6.92 (s, 1H), 6.64 (s, 1H), 4.18 (q, *J* = 7.1 Hz, 2H), 3.74 (m, 5H), 3.60 (dd, *J* = 14.0, 4.0 Hz, 1H), 3.53 – 3.39 (m, 2H), 3.04-2.91 (m, 3H), 2.59 (dd, *J* = 13.6, 4.3 Hz, 1H), 2.47 (m, 2H), 2.12 – 2.02 (m, 2H), 1.86 – 1.68 (m, 5H), 1.54 (s, 3H), 1.46 (s, 3H), 1.38 (t, *J* = 7.0 Hz, 3H), 1.28 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, MeOD) δ 173.92, 155.10, 154.26, 149.14, 147.89, 142.73, 140.50, 135.47, 133.15, 131.29, 131.07, 130.16, 128.95, 128.72, 125.96, 96.65, 95.36, 60.57, 59.00, 55.06, 50.45, 46.60, 41.84, 33.42, 30.25, 28.26, 27.77, 27.37, 25.08, 22.27, 20.89, 20.33, 13.65, 10.89. HRMS (FAB+) Calcd. for $C_{30}H_{40}N_3O_6S^+$ [M+1]: 570.2632; found 570.2629.



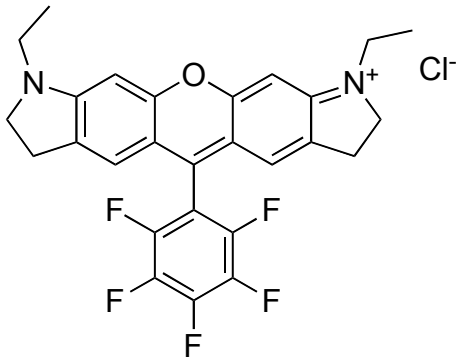
Xanthene 4-7a. Following the general procedure, diaryl ether **4-3d** (154 mg, 0.500 mmol) was reacted with benzoyl chloride (0.46 mL, 4.0 mmol) to provide 120 mg of **4-7a** as a deep purple residue (56%) and 55 mg of recovered **4-3d** (36%). ^{1}H NMR (400 MHz, MeOD) δ 7.70-7.65 (m, 3H), 7.45-7.40 (m, 2H), 6.96 (t, *J* = 1.7 Hz, 2H), 6.72 (s, 2H), 3.92 (t, *J* = 7.8 Hz, 4H), 3.60 (q, *J* = 7.2 Hz, 4H), 3.15 – 3.08 (m, 4H), 1.33 (t, *J* = 7.3 Hz, 6H). ^{13}C NMR (300 MHz, MeOD) δ 159.69, 159.47, 154.79, 134.33, 133.47, 129.89, 129.43, 129.00, 123.22, 114.41, 90.55, 51.93, 41.27, 26.08, 10.92. HRMS (FAB) Calcd. for $C_{27}H_{27}N_2O^+$ [M^+]: 395.2118; found 395.2122.



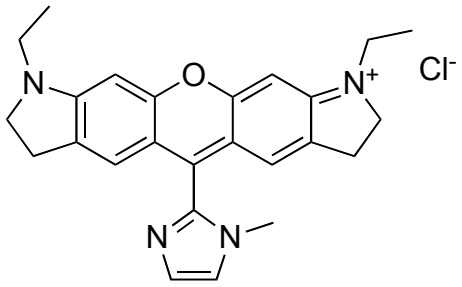
Xanthene 4-7c. Following the general procedure, diaryl ether **4-3d** (154 mg, 0.500 mmol) was reacted with *p*-(Me_2N)-benzoyl chloride (460 mg, 2.5 mmol) to provide 182 mg of **4-7c** as a deep red residue (77%). ^{1}H NMR (400 MHz, MeOD) δ 7.30 (d, *J* = 8.8 Hz, 2H), 7.19 (s, 2H), 7.02 (d, *J* = 8.8 Hz, 2H), 6.64 (s, 2H), 3.89 (t, *J* = 7.8 Hz, 4H), 3.57 (q, *J* = 7.2 Hz, 4H), 3.17 – 3.08 (m, 10H), 1.32 (t, *J* = 7.2 Hz, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.55, 159.12, 151.82, 133.76, 131.33, 123.76, 121.12, 120.45, 114.34, 112.31, 90.50, 51.82, 41.21, 39.67, 26.16, 10.91. HRMS (FAB) Calcd. for $C_{29}H_{32}N_3O^+$ [M^+]: 438.2540; found 438.2545.



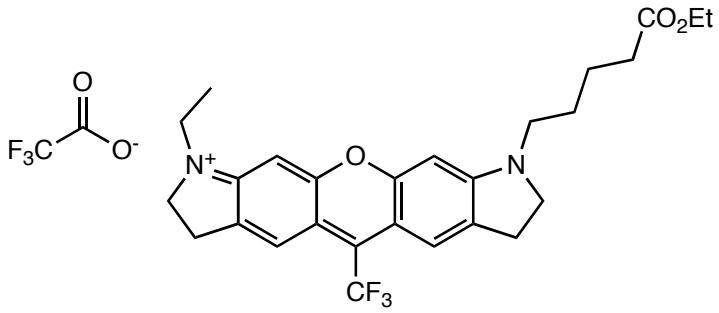
Xanthene 4-7d. Following the general procedure, diaryl ether **4-3d** (154 mg, 0.500 mmol) was reacted with *p*-fluorobenzoyl chloride (0.47 mL, 4.0 mmol) to provide 106 mg of **4-7d** as a deep purple residue (48%) and 63 mg of recovered **4-3d** (41%). ^{1}H NMR (400 MHz, MeOD) δ 7.50 – 7.39 (m, 4H), 6.96 (s, 2H), 6.71 (s, 2H), 3.92 (t, *J* = 7.8 Hz, 4H), 3.61 (q, *J* = 7.2 Hz, 4H), 3.19 – 3.08 (m, 4H), 1.33 (t, *J* = 7.2 Hz, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.65, 159.48, 153.65, 134.46, 131.84, 131.73, 129.45, 123.07, 116.24, 115.94, 114.50, 90.59, 51.95, 41.29, 26.10, 10.92. HRMS (FAB) Calcd. for $C_{27}H_{26}N_2FO^+$ [M^+]: 413.2024; found 413.2017.



Xanthene 4-7e. Following the general procedure, diaryl ether **4-3d** (154 mg, 0.500 mmol) was reacted with 2,3,4,5,6-pentafluorobenzoyl chloride (0.57 mL, 4.0 mmol) to provide 138 mg of **4-7e** as a green/blue solid (53%) and 61 mg of recovered **4-3d** (40%). ^{1}H NMR (400 MHz, MeOD) δ 7.06 (s, 2H), 6.77 (s, 2H), 3.97 (t, *J* = 7.6 Hz, 4H), 3.64 (q, *J* = 7.2 Hz, 4H), 3.27 – 3.10 (m, 4H), 1.35 (t, *J* = 7.3 Hz, 6H). ^{13}C NMR (300 MHz, MeOD) δ 159.87, 159.45, 145.99, 142.80, 140.39, 137.53, 135.81, 121.90, 114.53, 107.73, 90.91, 52.21, 41.46, 26.09, 10.97. HRMS (FAB) Calcd. for $C_{27}H_{22}N_2F_5O^+$ [M^+]: 485.1647; found 485.1663.



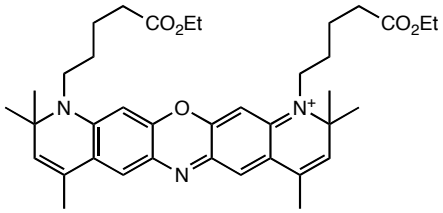
Xanthene 4-7f. Following the general procedure, diaryl ether **4-3d** (154 mg, 0.500 mmol) was reacted with 1-methyl-1*H*-imidazole-2-carbonyl chloride (361 mg, 2.5 mmol) to provide 78 mg of **4-7e** as a deep blue-purple solid (36%) and 68 mg of recovered **4-3d** (44%). ^{1}H NMR (400 MHz, MeOD) δ 7.95 (d, *J* = 1.9 Hz, 1H), 7.92 (d, *J* = 1.9 Hz, 1H), 6.85 – 6.81 (m, 4H), 4.01 (t, *J* = 7.5 Hz, 4H), 3.80 (s, 3H), 3.67 (q, *J* = 7.3 Hz, 4H), 3.26 – 3.18 (m, 4H), 1.35 (t, *J* = 7.3 Hz, 6H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 159.98, 159.36, 138.08, 136.83, 130.75, 125.95, 122.11, 120.92, 114.93, 91.53, 52.45, 41.63, 34.80, 26.12, 11.04. HRMS (FAB) Calcd. for $C_{27}H_{27}N_4O^+$ [M^+]: 399.2179; found 399.2197.

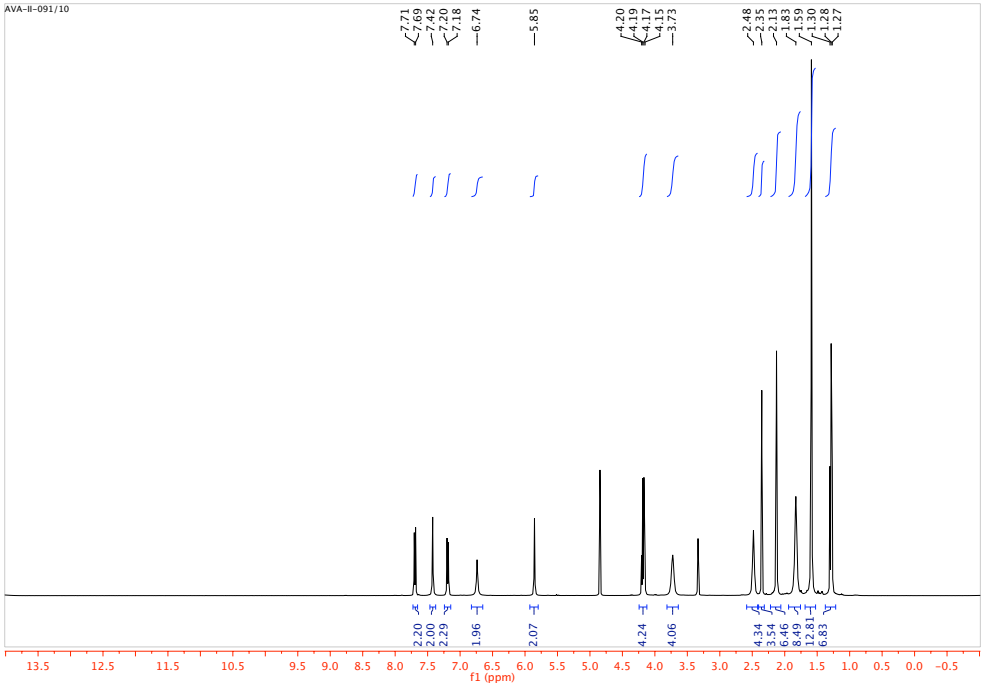


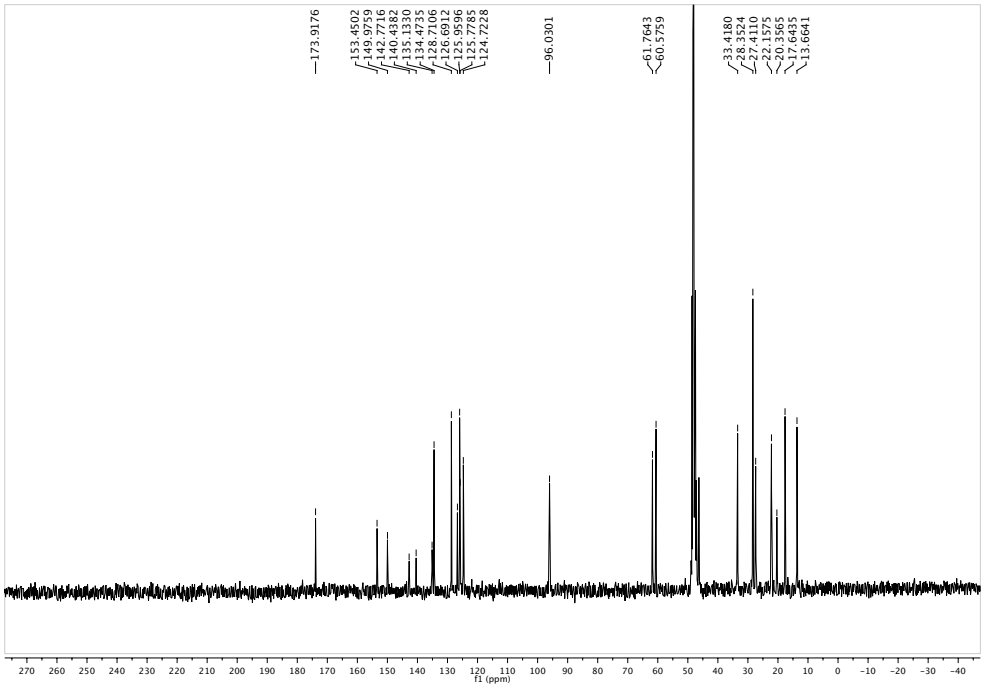
Xanthene 4-7b. Diaryl ether **4-3f** (1.15 g, 2.81 mmol) in anhydrous dichloromethane (30 mL) was treated dropwise via a pressure equalizing addition funnel with a solution of trifluoroacetic anhydride (0.87 mL, 6.2 mmol) in 20 mL of DCM over the course of approximately 2 hours. After addition was complete, the reaction was left to stir at room temperature overnight. The next day, the deep blue reaction mixture was quenched by the addition of 4% aqueous $NaHCO_3$, then extracted with dichloromethane (3 x 30 mL). The combined organic layers were dried over Na_2SO_4 and then concentrated under reduced pressure. The crude product was purified by flash chromatography (0-10% methanol/dichloromethane) to provide 1.40 g of **4-7b** as a green/blue solid (83%). ^{1}H NMR (400 MHz, MeOD) δ 7.72 (s, 2H), 6.76 (s, 1H), 6.74 (s, 1H), 4.15 (q, *J* = 7.1 Hz, 2H), 4.01 (td, *J* = 7.4, 3.7 Hz, 4H), 3.69 – 3.58 (m, 4H), 3.32-3.25 (m, 4H), 2.44 (t, *J* = 7.0 Hz, 2H), 1.86 – 1.68 (m, 4H), 1.35 (t, *J* = 7.3 Hz, 3H), 1.26 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 173.99, 159.95, 159.85, 159.51, 159.12, 136.61, 136.27, 120.92, 120.84, 120.77,

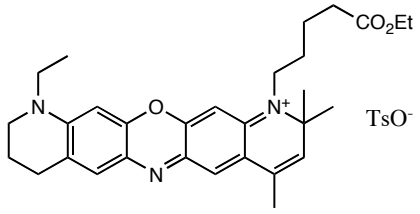
111.85, 111.67, 91.49, 91.33, 60.53, 52.87, 52.36, 46.61, 41.61, 33.43, 26.42, 22.23, 13.52, 11.09. HRMS (FAB) Calcd. For $C_{27}H_{30}F_3N_2O_3^+$ [M^+]: 487.2203; found 487.2205.

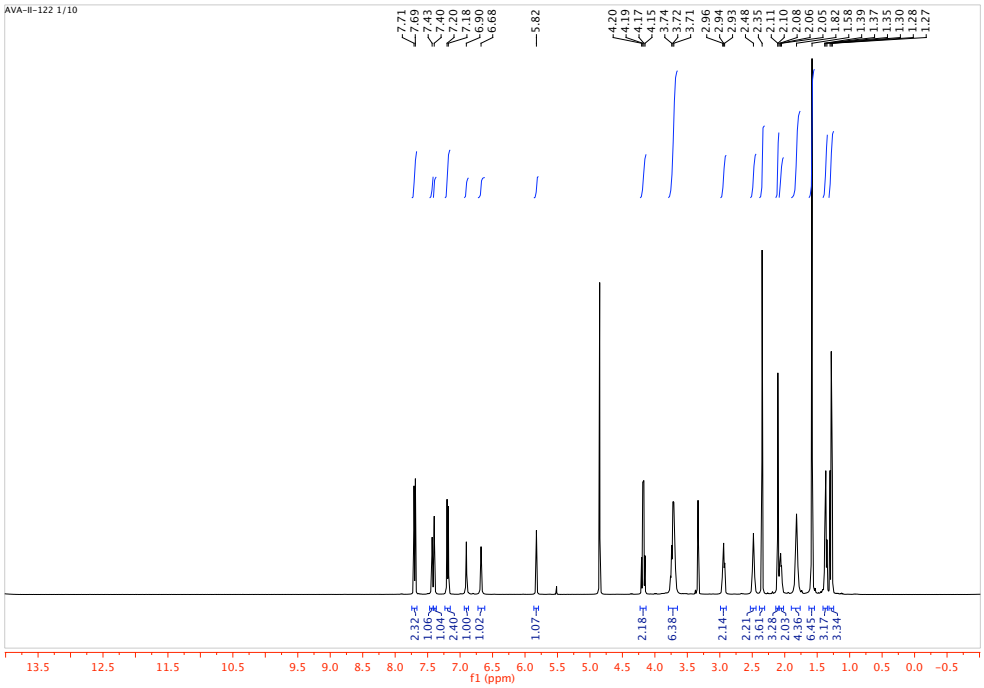
4.5 Spectra

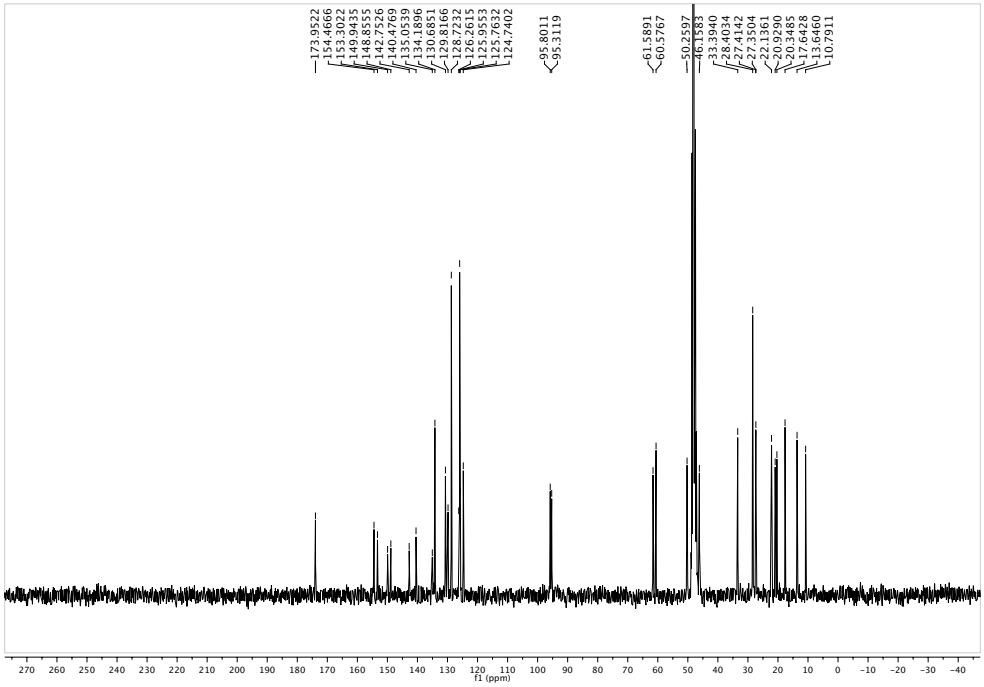

CO2Et
CO2Et
N
O
N+
N

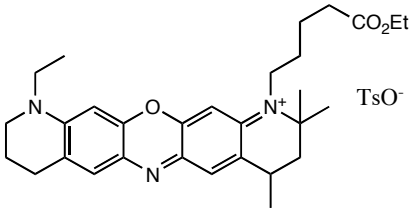

AVA-II-091/10
7.71
7.69
7.42
7.20
7.18
6.74
5.85
4.20
4.19
4.17
4.15
3.73
2.48
2.35
2.13
1.83
1.59
1.30
1.28
1.27
2.20
2.00
2.29
1.96
2.07
4.24
4.06
4.34
3.54
6.46
6.49
12.81
6.83
f1 (ppm)

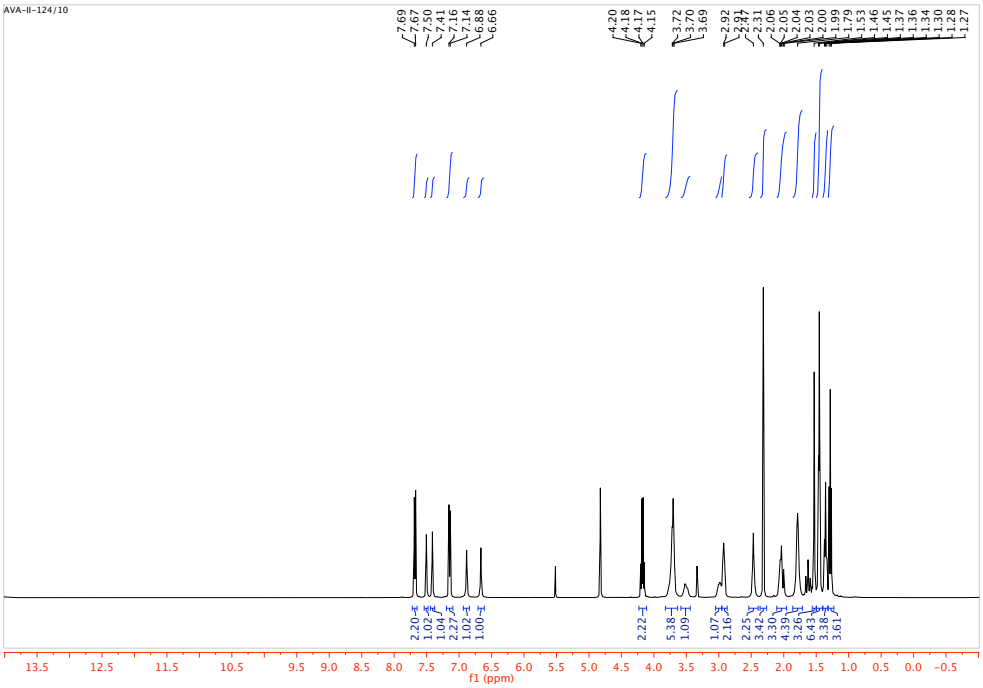

173.9176
153.4502
149.9759
142.7716
140.4382
135.1330
134.4735
128.7706
126.6912
125.9596
125.7785
124.7228
96.0301
61.7643
60.5759
33.4180
29.3154
27.4110
22.1575
20.3565
17.6435
13.6641
f1 (ppm)

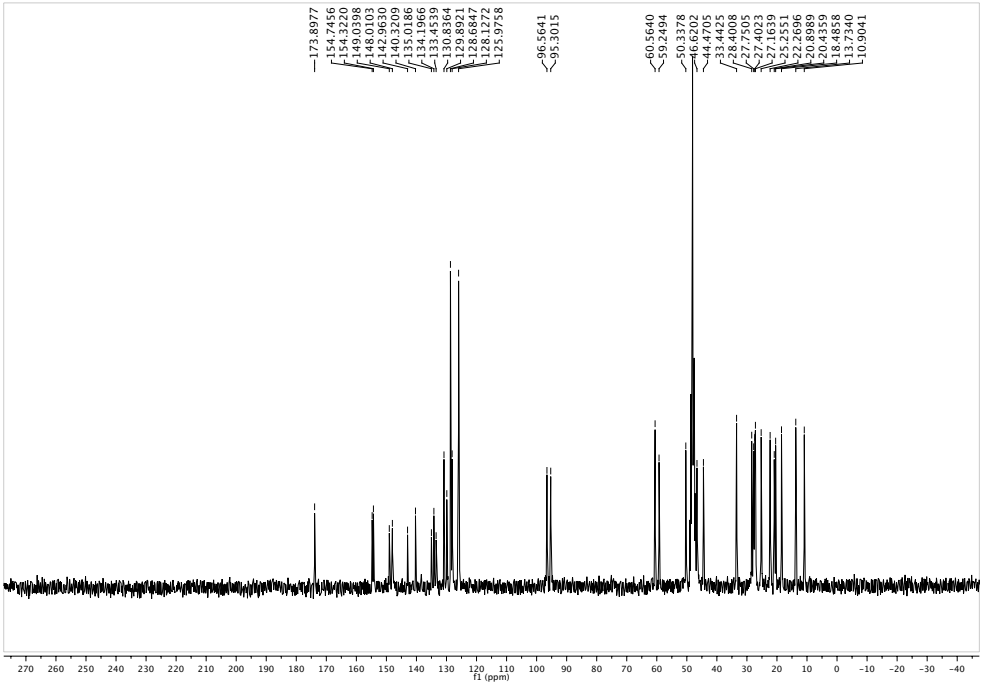

CO2Et
N
O
N+
TsO-
N

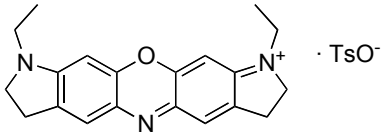

AVA-II-122 1/10
f1 (ppm)

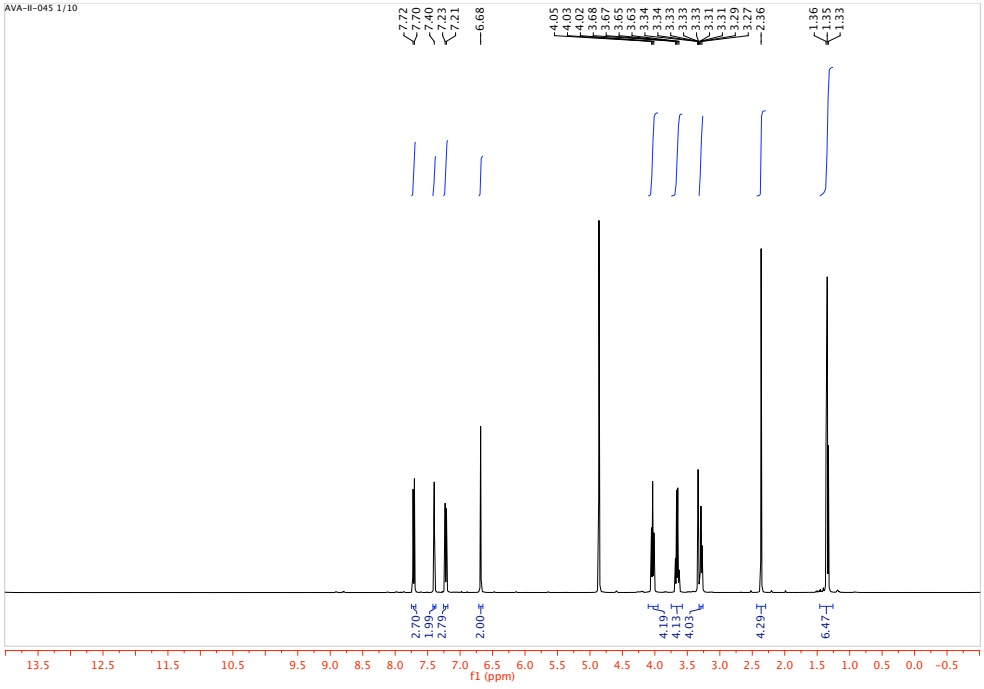

f1 (ppm)

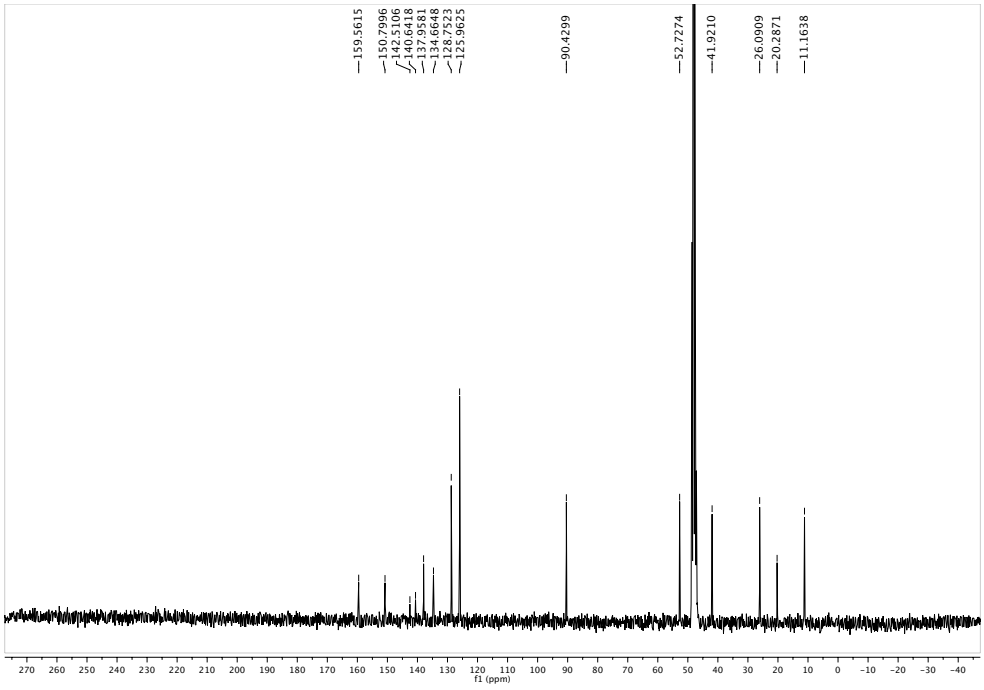

CO2Et
TsO-
N
O
N+
N

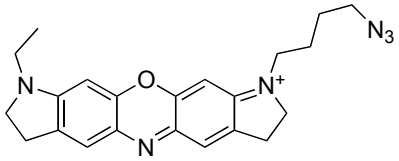

AVA-II-124/10
f1 (ppm)

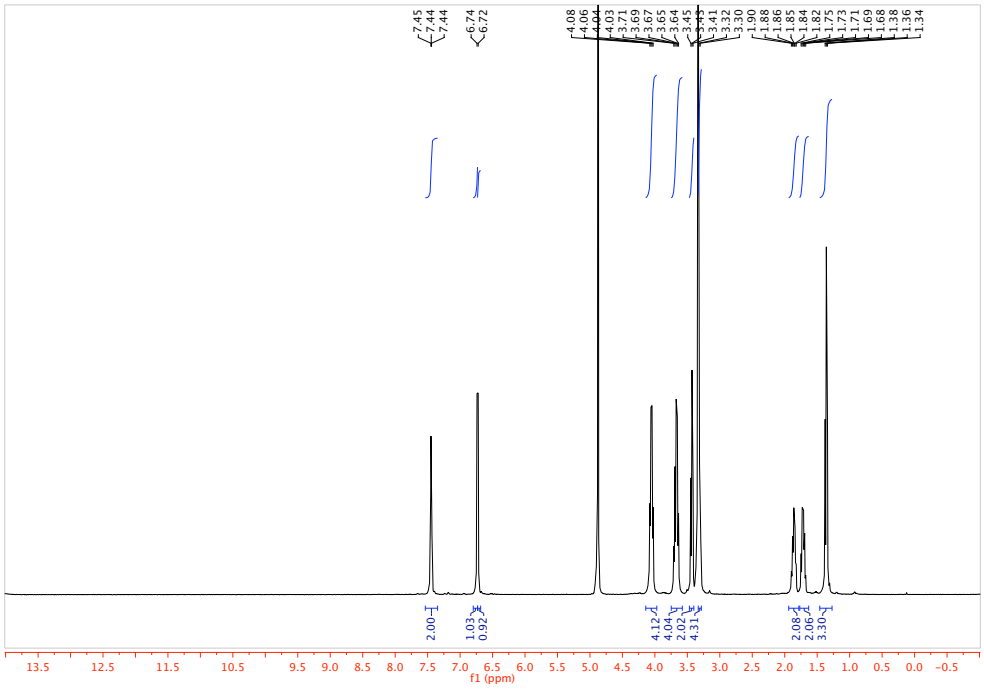

f1 (ppm)

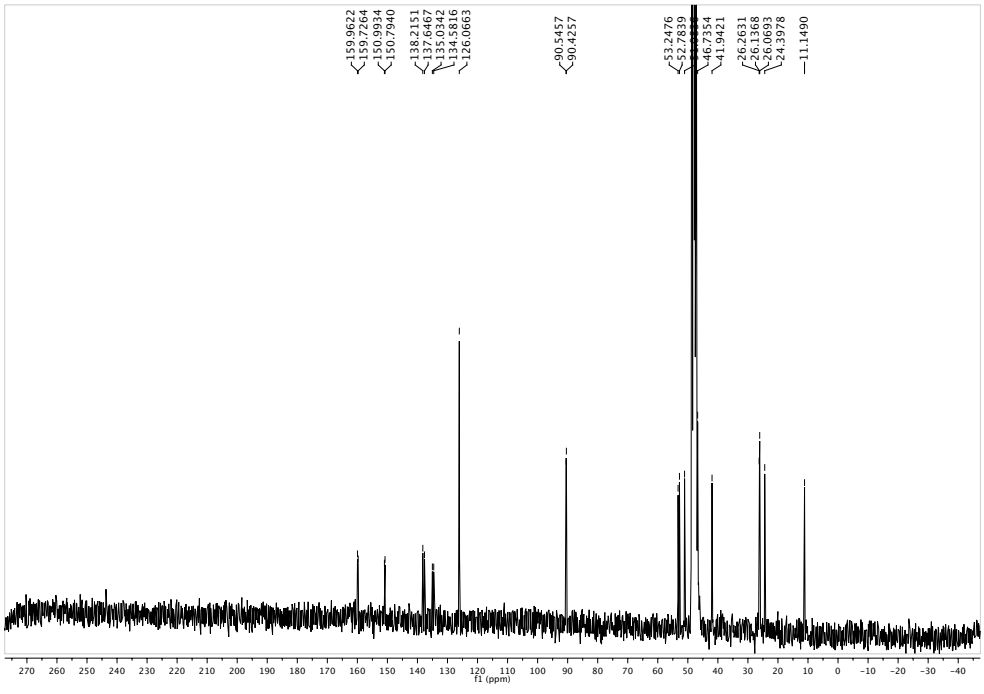

N
O
N+
N
· TsO-

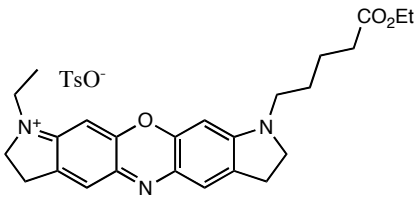

AVA-II-045 1/10
7.72
7.70
7.40
7.23
7.21
6.68
4.05
4.03
4.02
3.68
3.67
3.65
3.63
3.34
3.34
3.33
3.33
3.31
3.31
3.29
3.27
2.36
1.36
1.35
1.33
2.70
1.99
2.79
2.00
4.19
4.13
4.03
4.29
6.47
13.5 12.5 11.5 10.5 9.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0 -0.5
f1 (ppm)

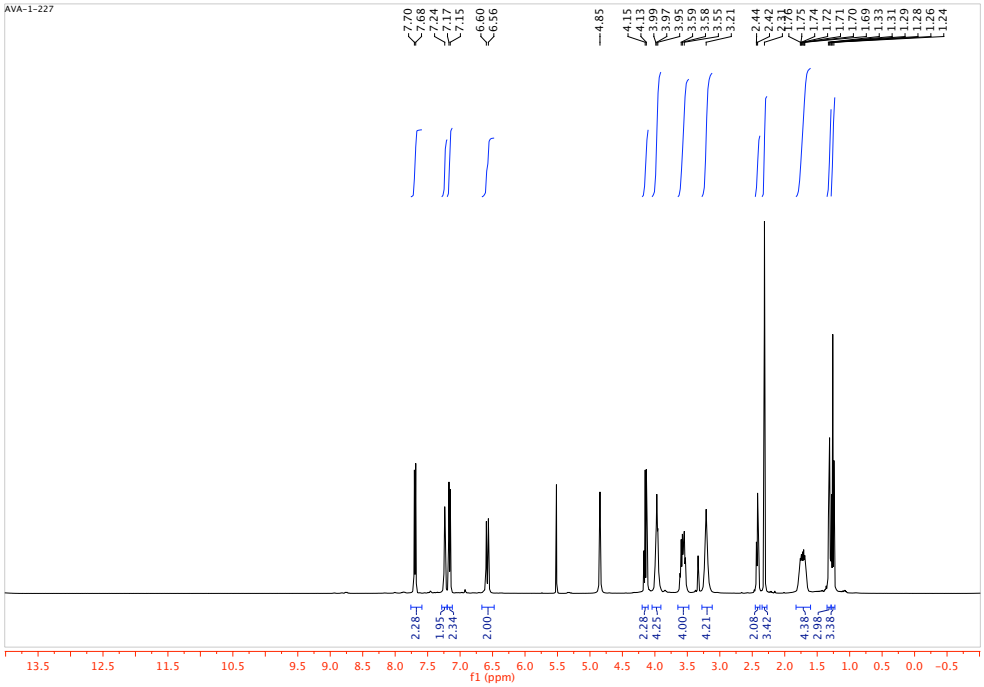

159.5615
150.7996
142.5106
140.6418
137.9581
134.6648
128.7523
125.9625
90.4299
52.7274
41.9210
26.0909
20.2871
11.1638
270 260 250 240 230 220 210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40
f1 (ppm)

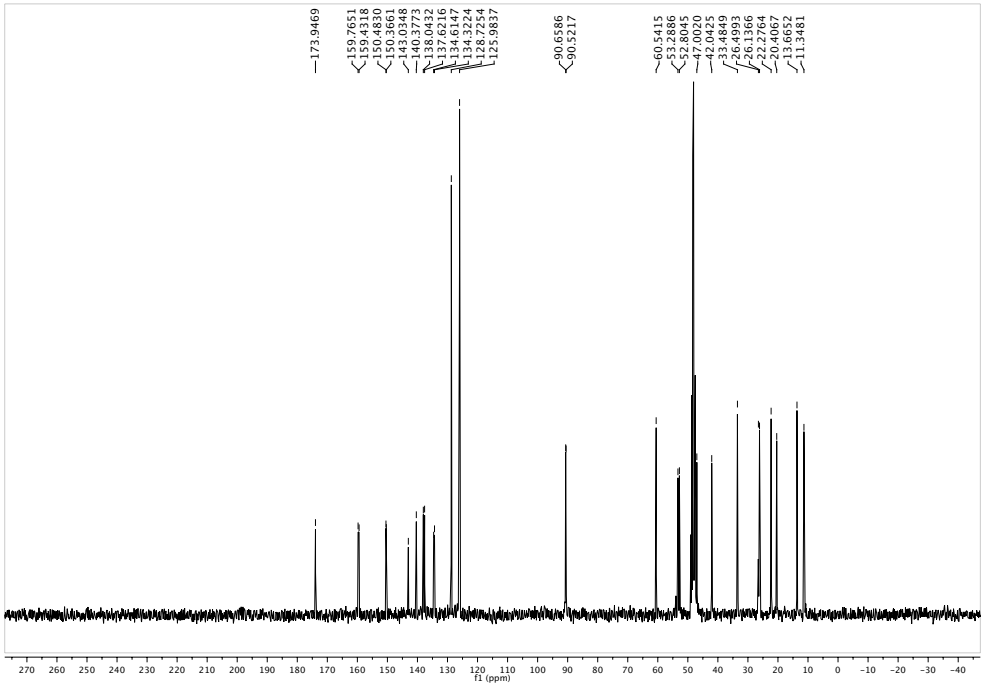

N3
N
O
N+
N

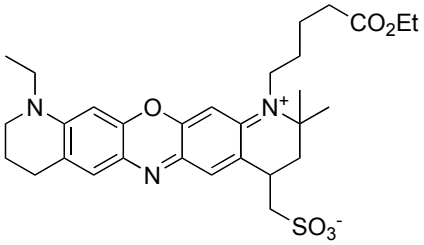

7.45
7.44
7.44
6.74
6.72
4.08
4.06
4.04
4.03
3.71
3.69
3.67
3.65
3.64
3.45
3.41
3.32
3.30
1.90
1.88
1.86
1.85
1.84
1.82
1.75
1.73
1.71
1.69
1.68
1.38
1.36
1.34
2.00
1.03
0.92
4.12
4.04
2.02
4.31
2.08
2.06
3.30
f1 (ppm)

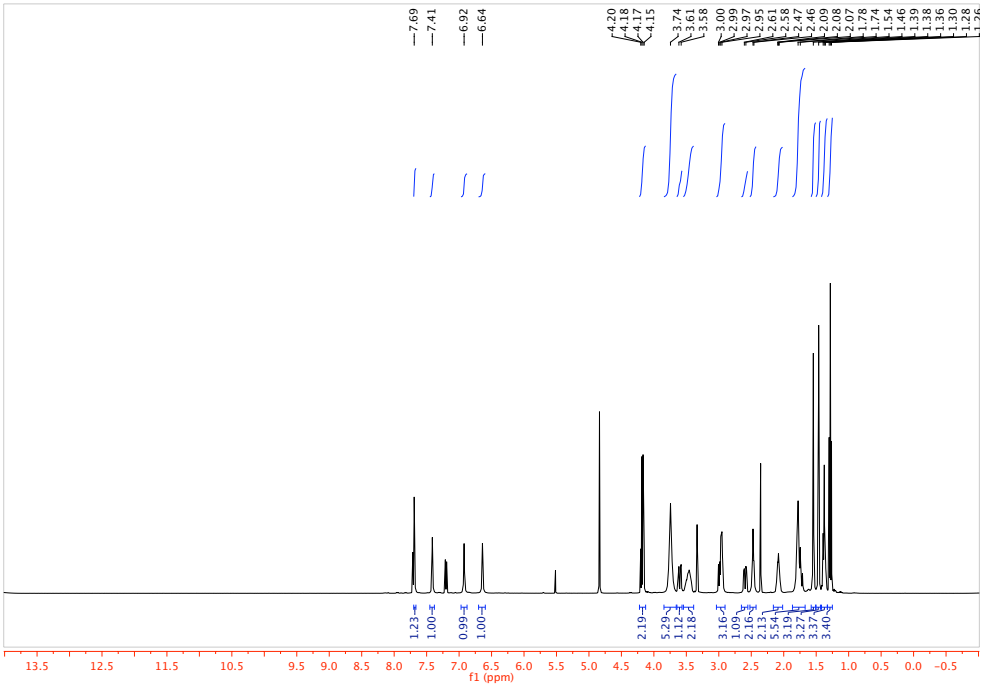

159.9622
159.7264
150.9934
150.7940
138.2151
137.6467
135.0342
134.5816
126.0663
90.5457
90.4257
53.2476
52.7839
46.7354
41.9421
26.2631
26.1388
26.0693
24.3978
11.1490
f1 (ppm)

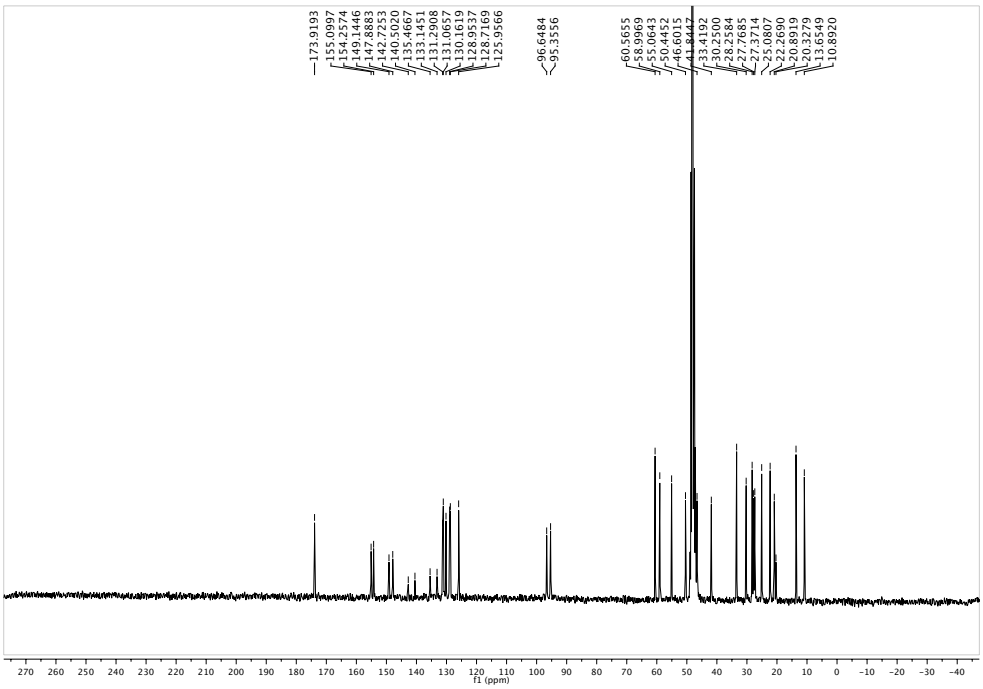

CO2Et
TsO-
N+
O
N
N

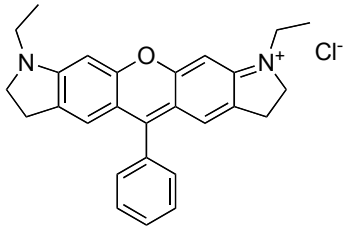

AVA-1-227
7.70
7.68
7.24
7.17
7.15
6.60
6.56
4.85
4.15
4.13
3.99
3.97
3.95
3.59
3.58
3.55
3.21
2.44
2.42
2.41
1.76
1.75
1.74
1.72
1.71
1.70
1.69
1.33
1.31
1.29
1.28
1.26
1.24
2.28
1.95
2.34
2.00
2.28
4.25
4.00
4.21
2.08
3.42
4.38
2.98
3.38
f1 (ppm)

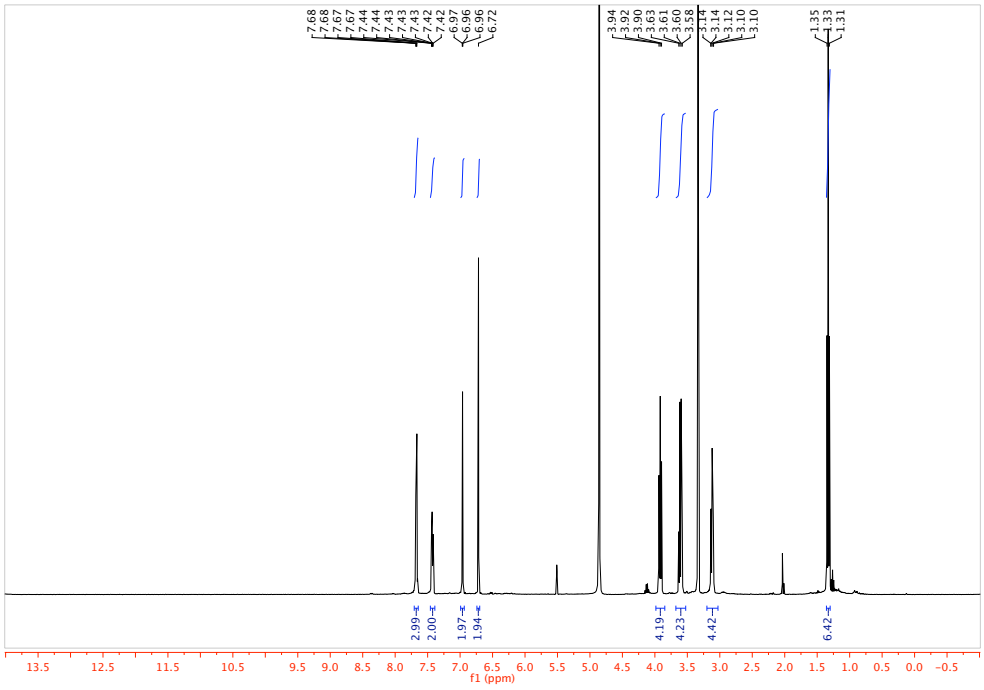

173.9469
159.7651
159.4318
158.8430
150.3661
143.0348
140.3773
138.0432
137.6216
134.6147
134.3224
128.7254
125.9837
90.6586
90.5217
60.5415
53.2846
52.8845
47.0020
42.0425
33.4849
26.4993
26.1366
22.2764
20.4067
13.6652
11.3481
f1 (ppm)

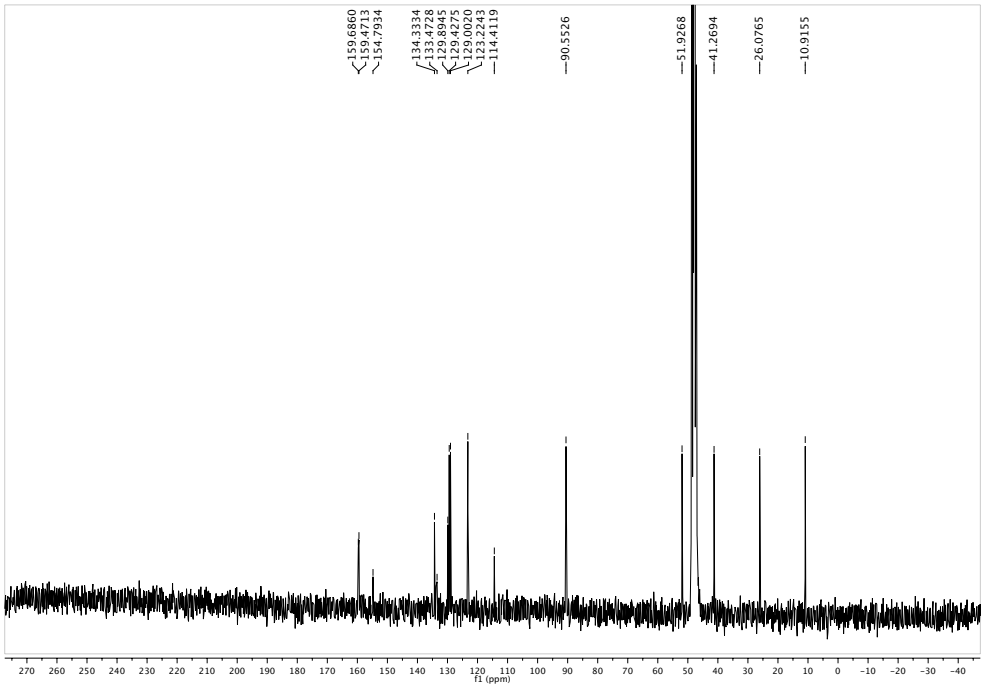

CO2Et
N
O
N+
N
SO3-

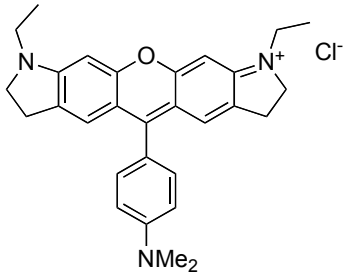

7.69
7.41
6.92
6.64
4.20
4.18
4.17
4.15
3.74
3.61
3.58
3.00
2.99
2.97
2.95
2.61
2.58
2.47
2.46
2.09
2.08
2.07
1.78
1.74
1.54
1.46
1.39
1.38
1.36
1.30
1.28
1.23
1.00
0.99
1.00
2.19
5.29
1.12
2.18
3.16
1.09
2.16
2.13
5.54
3.19
3.27
3.37
3.40
f1 (ppm)

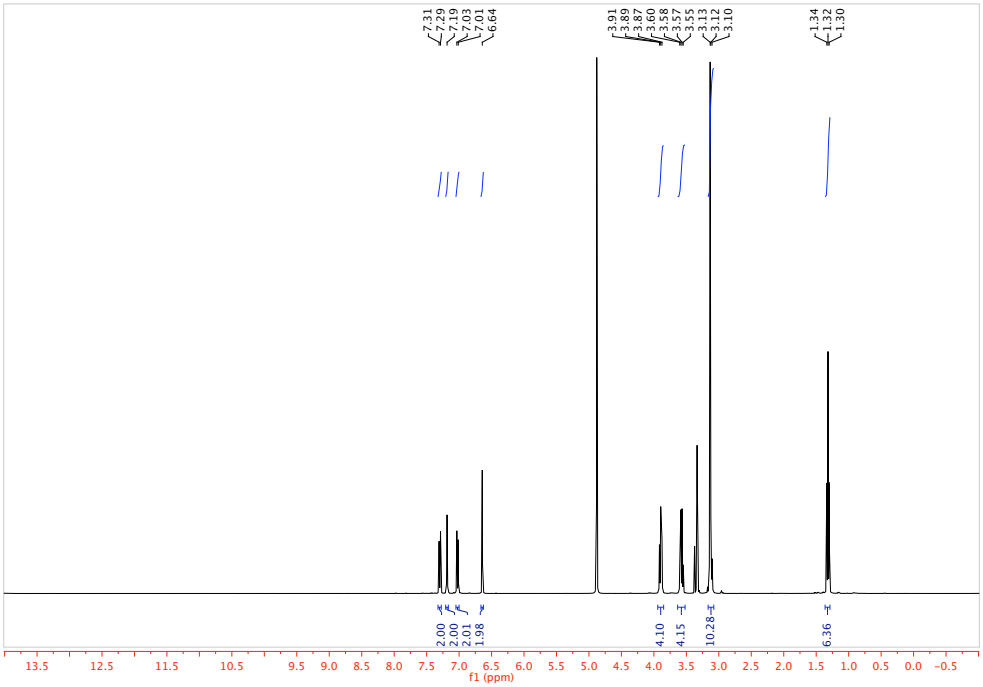

173.9193
155.0997
154.2574
149.1446
147.8883
142.7253
140.5020
135.4667
133.1451
131.2908
131.0657
130.1519
128.9537
128.7169
125.9566
96.6484
95.3556
60.5655
58.9969
55.0643
50.4452
46.6015
41.8447
33.4192
30.2500
28.2584
27.7685
27.3714
25.5087
22.2690
20.8919
20.3279
13.6549
10.8920
f1 (ppm)

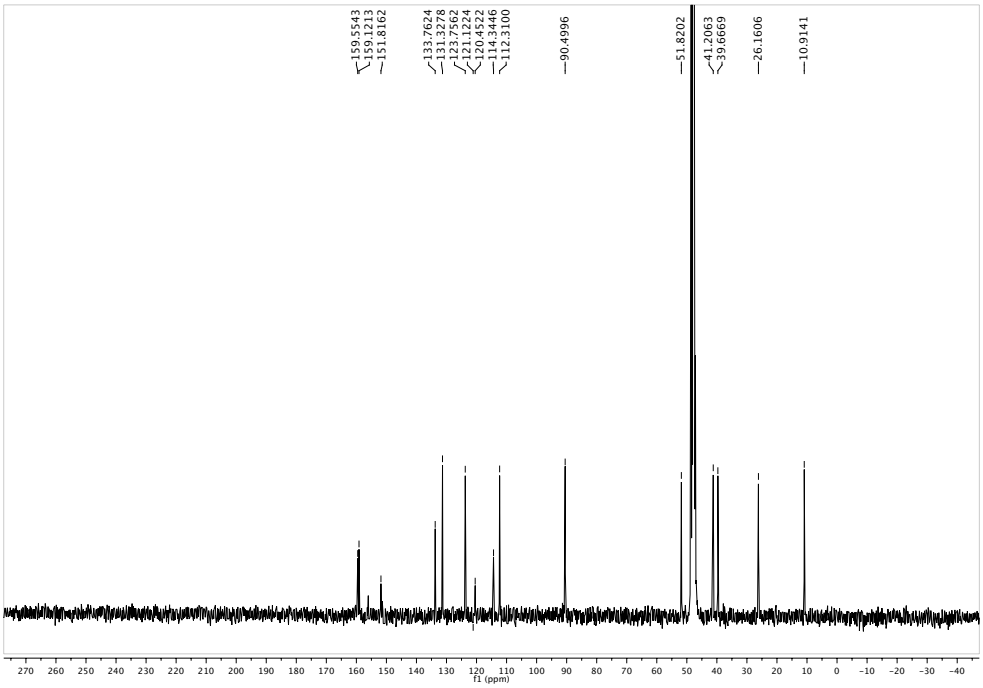

N
O
N+
Cl-

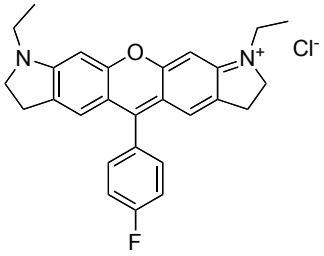

7.68
7.68
7.67
7.67
7.44
7.44
7.43
7.43
7.42
7.42
6.97
6.96
6.96
6.72
3.94
3.92
3.90
3.63
3.61
3.60
3.58
3.14
3.12
3.10
3.10
1.35
1.33
1.31
2.99
2.00
1.97
1.94
4.19
4.23
4.42
6.42
f1 (ppm)

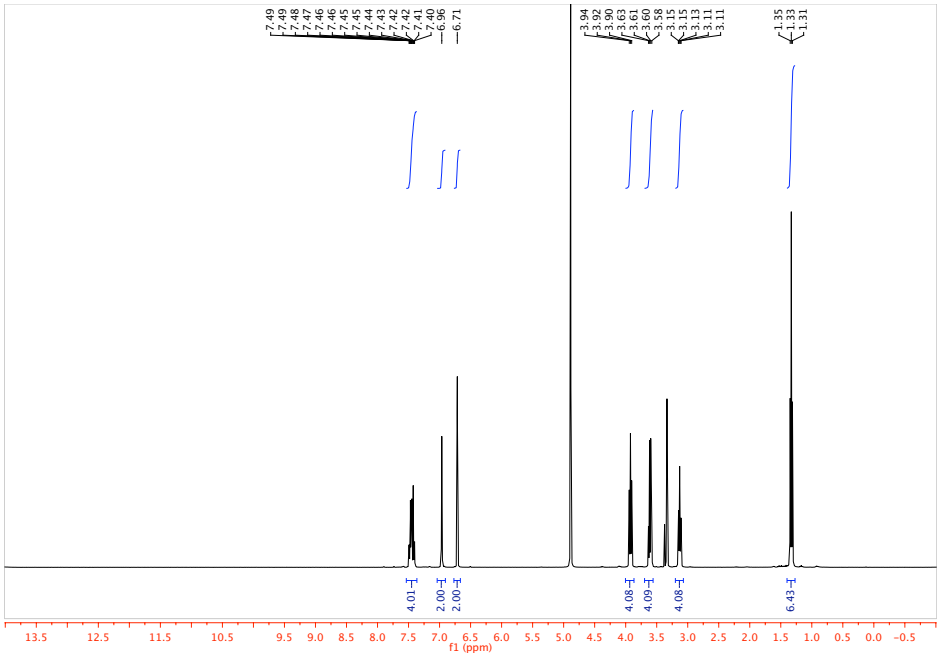

159.6860
159.4713
154.7934
134.3334
133.4728
129.8945
129.4275
129.0020
123.2243
114.4119
90.5526
51.9268
41.2694
26.0765
10.9155
f1 (ppm)

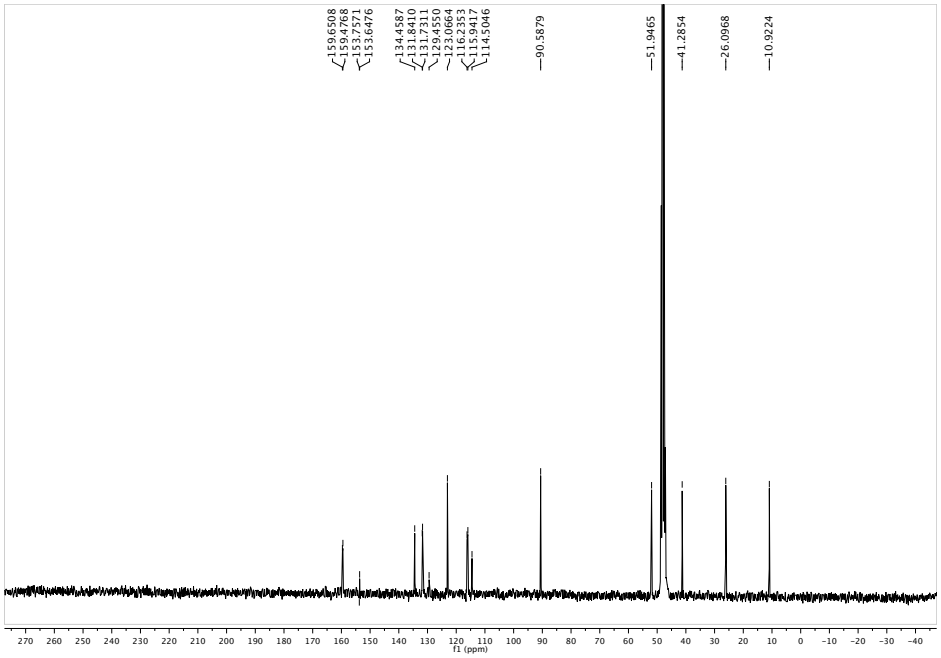

N
O
N+
Cl-
NMe2

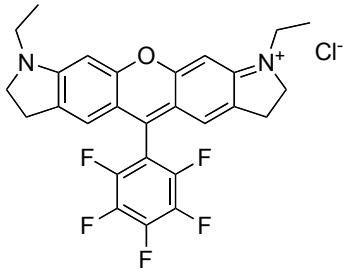

7.31
7.29
7.19
7.03
7.01
6.64
3.91
3.89
3.87
3.60
3.58
3.57
3.55
3.13
3.12
3.10
1.34
1.32
1.30
2.00
2.00
2.01
1.98
4.10
4.15
10.28
6.36
f1 (ppm)

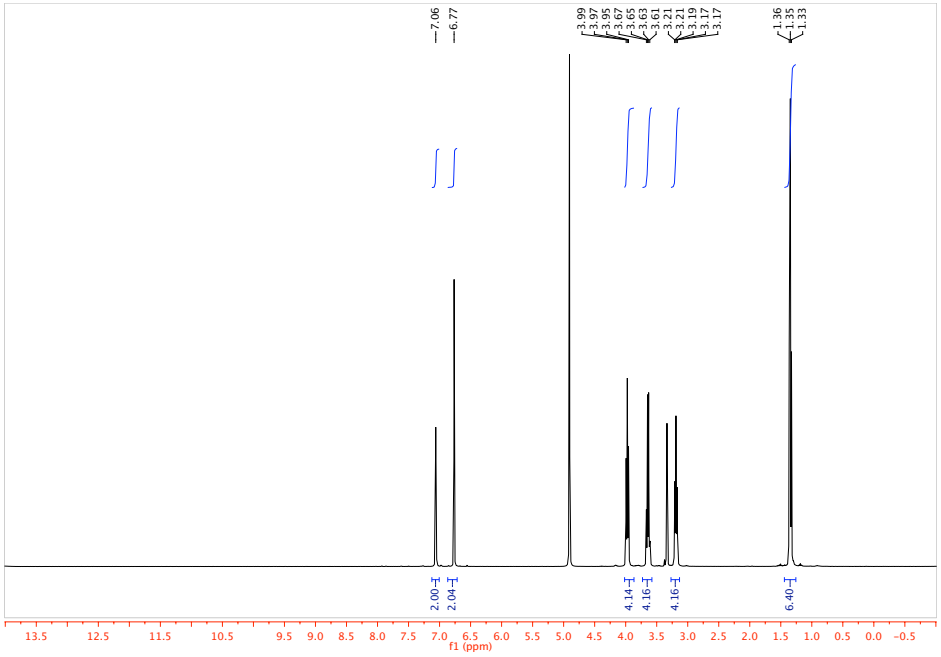

159.5543
159.1213
151.8162
133.7624
131.3278
123.7562
121.1224
120.4422
114.3446
112.3100
90.4996
51.8202
41.2063
39.6669
26.1606
10.9141
f1 (ppm)

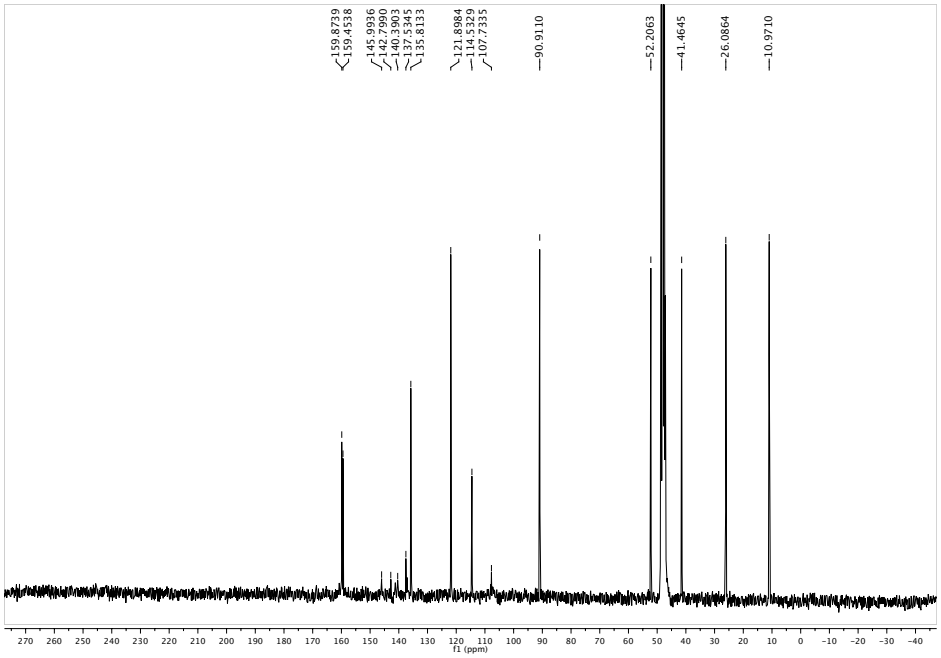

N
O
N+
Cl-
F

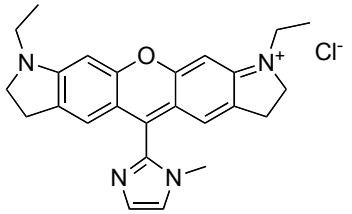

7.49
7.49
7.48
7.47
7.46
7.46
7.45
7.45
7.44
7.43
7.42
7.42
7.41
7.40
6.96
6.71
3.94
3.92
3.90
3.63
3.61
3.60
3.58
3.15
3.15
3.13
3.11
3.11
1.35
1.33
1.31
4.01
2.00
2.00
4.08
4.09
4.08
6.43
f1 (ppm)

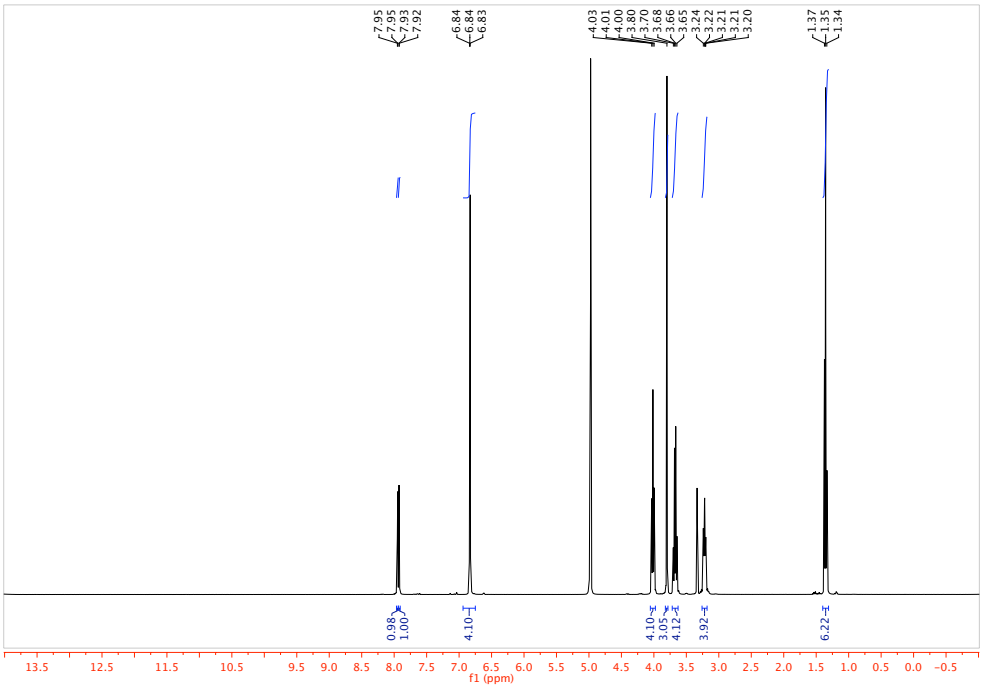

159.6508
159.4768
153.7571
153.6476
134.4587
131.8410
131.7211
129.4550
123.0664
116.2353
115.9417
114.5046
90.5879
51.9465
41.2854
26.0968
10.9224
f1 (ppm)

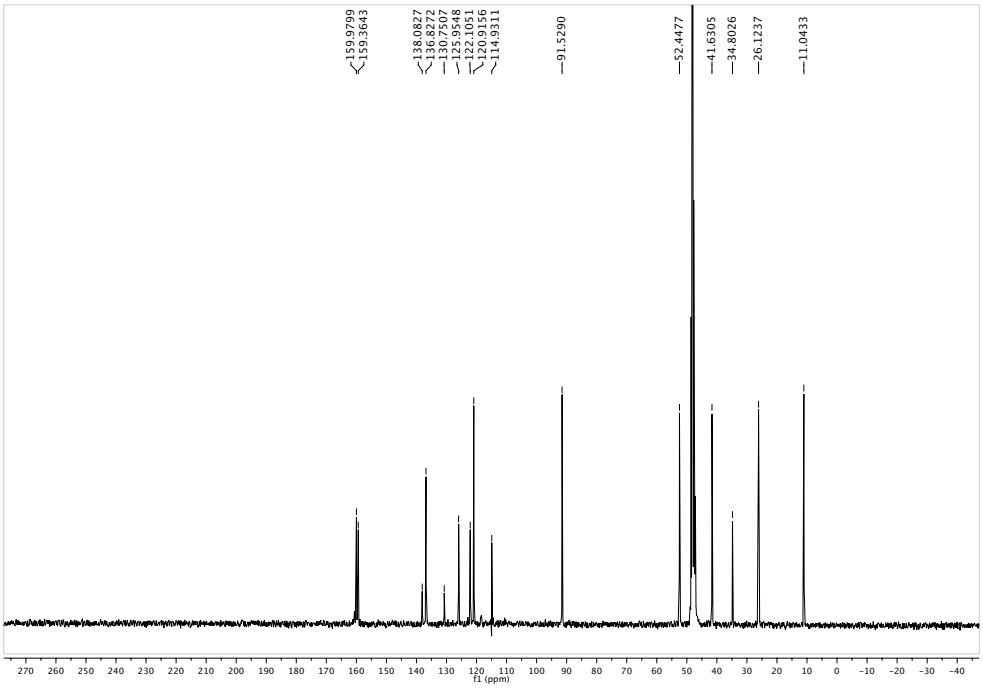

N
O
N+
Cl-
F
F
F
F
F

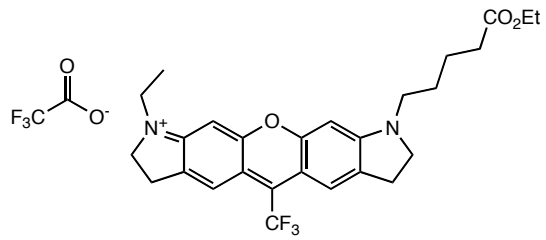

7.06
6.77
3.99
3.97
3.95
3.67
3.65
3.63
3.61
3.21
3.19
3.17
3.17
1.36
1.35
1.33
2.00
2.04
4.14
4.16
4.16
6.40
f1 (ppm)

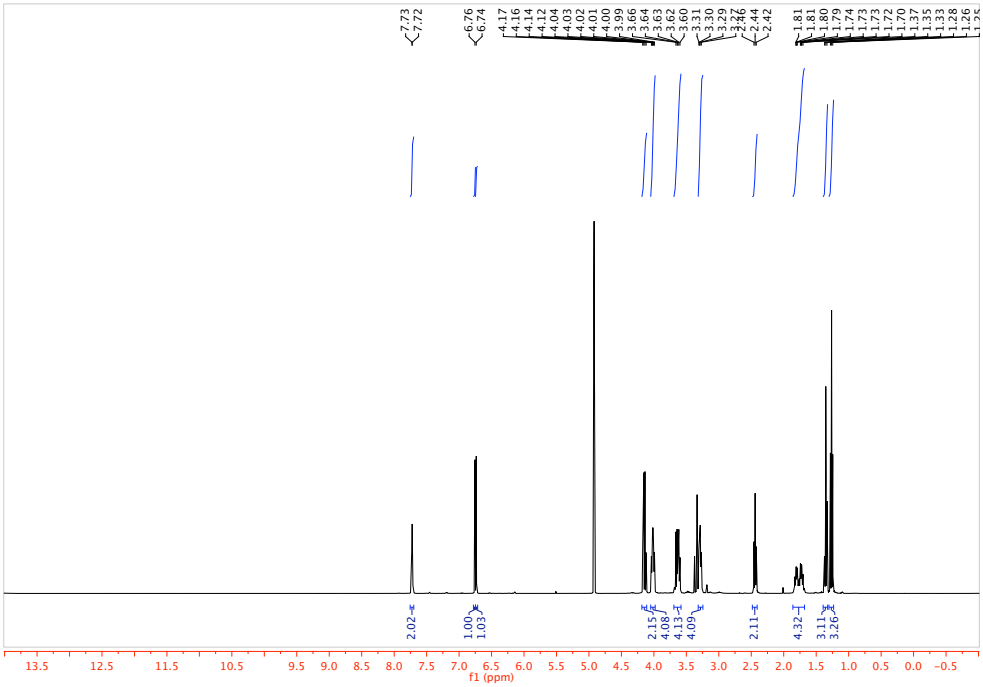

159.8739
159.4538
145.9936
142.7990
140.3903
137.5345
135.8133
121.8984
114.5329
107.7335
90.9110
52.2063
41.4645
26.0864
10.9710
f1 (ppm)

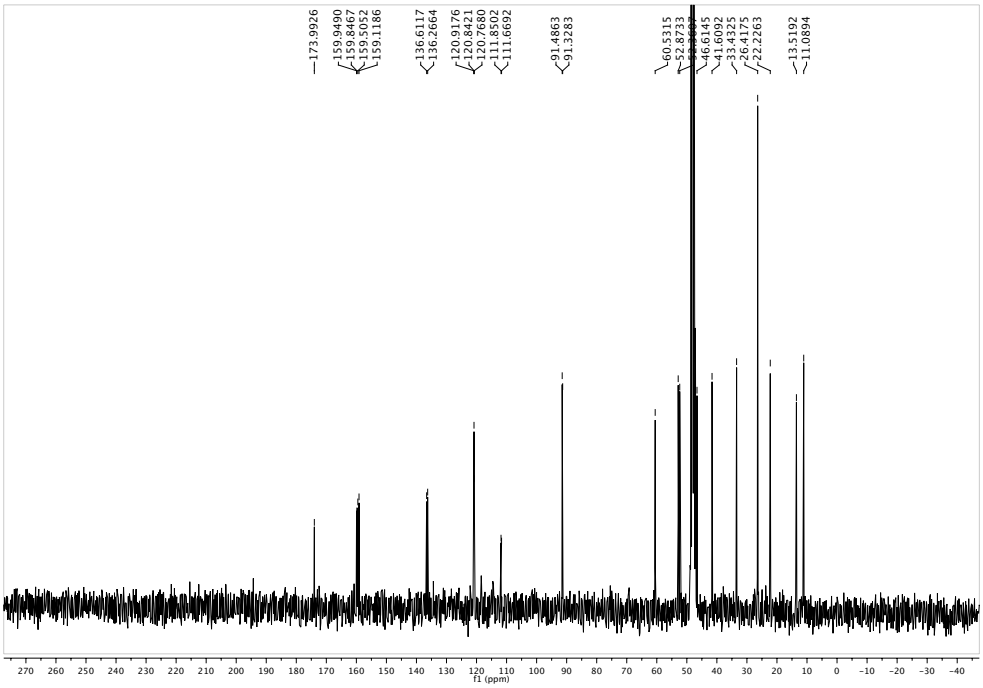

N
O
N+
Cl-
N
N


7.95
7.95
7.93
7.92
6.84
6.84
6.83
4.03
4.01
4.00
3.80
3.70
3.68
3.66
3.65
3.24
3.22
3.21
3.21
3.20
1.37
1.35
1.34
0.98
1.00
4.10
4.10
3.05
4.12
3.92
6.22
f1 (ppm)


159.9799
159.3643
138.0827
136.8272
130.7507
125.9548
122.1051
120.9156
114.9311
91.5290
52.4477
41.6305
34.8026
26.1237
11.0433
f1 (ppm)


CO2Et
O
F3C
O-
N+
O
N
CF3


7.73
7.72
6.76
6.74
4.17
4.16
4.14
4.12
4.04
4.03
4.02
4.01
4.00
3.99
3.66
3.64
3.63
3.62
3.60
3.31
3.30
3.29
2.46
2.44
2.42
1.81
1.80
1.79
1.74
1.73
1.72
1.70
1.37
1.35
1.33
1.28
1.26
2.02
1.00
1.03
2.15
4.08
4.13
4.09
2.11
4.32
3.11
3.26
f1 (ppm)


173.9926
159.9490
159.8467
159.5052
159.1186
136.6117
136.2664
120.9176
120.8421
120.7680
111.8502
111.6692
91.4863
91.3283
60.5315
52.8733
46.6145
41.6092
33.4325
26.4175
22.2263
13.5192
11.0894
f1 (ppm)

4.6 References

1. Tinnefeld, P. & Sauer, M. Branching Out of Single-Molecule Fluorescence Spectroscopy: Challenges for Chemistry and Influence on Biology. *Angew. Chem. Int. Ed.* **44,** 2642–2671 (2005).
2. Huang, B., Bates, M. & Zhuang, X. Super-Resolution Fluorescence Microscopy. *Annu. Rev. Biochem.* **78,** 993–1016 (2009).
3. Fernández-Suárez, M. & Ting, A. Y. Fluorescent probes for super-resolution imaging in living cells. *Nat. Rev. Mol. Cell Biol.* **9,** 929–943 (2008).
4. Chen, Z., Cornish, V. W. & Min, W. Chemical tags: inspiration for advanced imaging techniques. *Curr. Opin. Chem. Biol.* **17,** 637–643 (2013).
5. Jing, C. & Cornish, V. W. Chemical Tags for Labeling Proteins Inside Living Cells. *Acc. Chem. Res.* **44,** 784–792 (2011).
6. Haidekker, M. A., Brady, T. P., Lichlyter, D. & Theodorakis, E. A. A Ratiometric Fluorescent Viscosity Sensor. *J. Am. Chem. Soc.* **128,** 398–399 (2006).
7. Charier, S. *et al.* An Efficient Fluorescent Probe for Ratiometric pH Measurements in Aqueous Solutions. *Angew. Chem. Int. Ed.* **43,** 4785–4788 (2004).
8. Grynkiewicz, G., Poenie, M. & Tsien, R. Y. A new generation of Ca2+ indicators with greatly improved fluorescence properties. *J. Biol. Chem.* **260,** 3440–3450 (1985).
9. Yu, F. *et al.* A Near-IR Reversible Fluorescent Probe Modulated by Selenium for Monitoring Peroxynitrite and Imaging in Living Cells. *J. Am. Chem. Soc.* **133,** 11030–11033 (2011).
10. Fluhler, E., Burnham, V. G. & Loew, L. M. Spectra, membrane binding, and potentiometric responses of new charge shift probes. *Biochemistry (Mosc.)* **24,** 5749–5755 (1985).
11. Dempsey, G. T., Vaughan, J. C., Chen, K. H., Bates, M. & Zhuang, X. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods* **8,** 1027–1036 (2011).
12. Fölling, J. *et al.* Photochromic Rhodamines Provide Nanoscopy with Optical Sectioning. *Angew. Chem. Int. Ed.* **46,** 6266–6270 (2007).

13. Heilemann, M., van de Linde, S., Mukherjee, A. & Sauer, M. Super-Resolution Imaging with Small Organic Fluorophores. *Angew. Chem. Int. Ed.* **48,** 6903–6908 (2009).
14. Wombacher, R. *et al.* Live-cell super-resolution imaging with trimethoprim conjugates. *Nat. Methods* **7,** 717–719 (2010).
15. Beija, M., Afonso, C. A. M. & Martinho, J. M. G. Synthesis and applications of Rhodamine derivatives as fluorescent probes. *Chem. Soc. Rev.* **38,** 2410–2433 (2009).
16. Kanitz, A. & Hartmann, H. Preparation and Characterization of Bridged Naphthoxazinium Salts. *Eur. J. Org. Chem.* **1999,** 923–930 (1999).
17. Pauff, S. M. & Miller, S. C. Synthesis of Near-IR Fluorescent Oxazine Dyes with Esterase-Labile Sulfonate Esters. *Org. Lett.* **13,** 6196–6199 (2011).
18. Gribble, G. W. *et al.* Reactions of sodium borohydride in acidic media. I. Reduction of indoles and alkylation of aromatic amines with carboxylic acids. *J. Am. Chem. Soc.* **96,** 7812–7814 (1974).
19. Ullmann, F. & Sponagel, P. Ueber die Phenylirung von Phenolen. *Berichte Dtsch. Chem. Ges.* **38,** 2211–2212 (1905).
20. Burgos, C. H., Barder, T. E., Huang, X. & Buchwald, S. L. Significantly Improved Method for the Pd-Catalyzed Coupling of Phenols with Aryl Halides: Understanding Ligand Effects. *Angew. Chem. Int. Ed.* **45,** 4321–4326 (2006).
21. Maiti, D. & Buchwald, S. L. Orthogonal Cu- and Pd-Based Catalyst Systems for the O- and N-Arylation of Aminophenols. *J. Am. Chem. Soc.* **131,** 17423–17429 (2009).
22. Klapars, A. & Buchwald, S. L. Copper-Catalyzed Halogen Exchange in Aryl Halides: An Aromatic Finkelstein Reaction. *J. Am. Chem. Soc.* **124,** 14844–14845 (2002).
23. Mitronova, G. Y. *et al.* New Fluorinated Rhodamines for Optical Microscopy and Nanoscopy. *Chem. – Eur. J.* **16,** 4477–4488 (2010).
24. Olah, G. A., Farooq, O., Farnia, S. M. F. & Olah, J. A. Friedel-Crafts chemistry. 11. Boron, aluminum, and gallium tris(trifluoromethanesulfonate) (triflate): effective new Friedel-Crafts catalysts. *J. Am. Chem. Soc.* **110,** 2560–2565 (1988).

25. Kobayashi, S., Komoto, I. & Matsuo, J. Catalytic Friedel-Crafts Acylation of Aniline Derivatives. *Adv. Synth. Catal.* **343,** 71–74 (2001).

26. Prakash, G. K. S., Mathew, T. & Olah, G. A. Gallium(III) Triflate: An Efficient and a Sustainable Lewis Acid Catalyst for Organic Synthetic Transformations. *Acc. Chem. Res.* **45,** 565–577 (2012).

27. Clunas, S. *et al.* 3,6-disubstituted xanthylium salts as medicaments. (2010). at <http://www.google.com/patents/WO2010067078A3>

28. *Principles of Fluorescence Spectroscopy*. (Springer US, 2006). at <http://link.springer.com/10.1007/978-0-387-46312-4>

29. Williams, A. T. R., Winfield, S. A. & Miller, J. N. Relative fluorescence quantum yields using a computer-controlled luminescence spectrometer. *Analyst* **108,** 1067–1071 (1983).

30. Yamada, Y. *et al.* Synthesis of Linear Tripeptides for Right-Hand Segments of Complestatin. *Chem. Pharm. Bull. (Tokyo)* **53,** 1277–1290 (2005).

31. Lowery, C. A., Park, J., Kaufmann, G. F. & Janda, K. D. An Unexpected Switch in the Modulation of AI-2-Based Quorum Sensing Discovered through Synthetic 4,5-Dihydroxy-2,3-pentanedione Analogues. *J. Am. Chem. Soc.* **130,** 9200–9201 (2008).

Chapter 5

A Photoactivatable Oxazine Fluorophore

for Live-cell Imaging

---

*The contents of this chapter will be published in:

A.V. Anzalone, Z. Chen, V.W. Cornish, et al "Design and synthesis of photoactivatable oxazine fluorophores for live-cell imaging." *In preparation*.

5.0 Chapter Outlook

Emerging technologies in optical imaging are providing methods for visualizing biomolecules at unprecedented resolution. As advances are made in the microscopy arena, the properties of fluorescent probes must evolve to meet the demands of new technologies. Recent developments in the field of super-resolution imaging have enabled sub-diffraction localization of individual molecules inside of living cells. These technologies require fluorophores with special properties such as photoswitching and photoactivatability. In addition, the resolution of the image depends on the total number of photons that can be detected per molecules, and therefore benefits from fluorophores with high photon outputs. Here, we report the design and synthesis of a photoactivatable azido-acyl oxazine fluorophore for live cell imaging that can be applied to meet these needs. Photoactivation is achieved cleanly and rapidly with 365 nm light, producing a single fluorescent oxazine photoproduct. We demonstrate protein specific labeling in living cells using trimethoprim conjugates, confirming the utility of azido-acyl oxazines for live cell imaging applications. Notably, the photoactivatable oxazine fluorophores outperform their standard oxazine counterparts, establishing photocaging as a feasible approach for overcoming challenges of cell-permeability and non-specific staining. The photoactivatable oxazine developed here provides a powerful new reagent for biological imaging, with potential application to super-resolution imaging and single molecule biophysical studies.

5.1 Introduction

With the continuing development of fluorescence-based tools for studying molecular processes within living cells comes the demand for brighter, multifunctional fluorescent probes[1]. Great strides have been made over the past several decades in devising chemical tools for protein specific labeling[2–7] and intracellular molecular sensing[8] using fluorescent reporters. Likewise, there has been significant progress in the development of single molecule super-resolution imaging technologies[9,10] for localizing biomolecules in cells with unprecedented precision. For live-cell imaging, these methods require probes with high overall brightness, high photostability, preferably with absorbance and emission of red or near-IR light, cellular permeability, and good cellular properties (i.e. devoid of non-specific staining)[1]. Several other advantageous properties, such as fluorogenicity[11] and photoactivatability[12], are also exploited in various applications. Thus, while the microscopy toolkit expands to meet new challenges in the imaging arena, so too should our repertoire of small molecule fluorophores in order to maximize the potential of emerging technologies.

Oxazines comprise a class of live-cell fluorescent imaging reagents exemplified by the popular commercial dye Atto 655. Their remarkable photostability, notably higher in comparison to the cyanine fluorophores, and emission in the red/far-red region of the visible spectrum, notably redder than the rhodamine dye class, make these fluorophores well-suited to live-cell and single molecule-based techniques. Oxazines also undergo photoswitching[13], which has been exploited for super-resolution dSTORM imaging[14]. Like many other classes of fluorophores, one drawback for oxazines is their unreliable behavior within the context of living cells, either due to low permeability or aggregation

leading to high background staining. Also, to date, the derivatization of oxazine scaffolds has not been extensively explored, potentially due to synthetic challenges. Therefore, the oxazine class of molecules would benefit from new synthetic methods for their assembly that enable the diversification of the scaffold with new chemical features, either to improve cellular behavior or to install useful functionality.

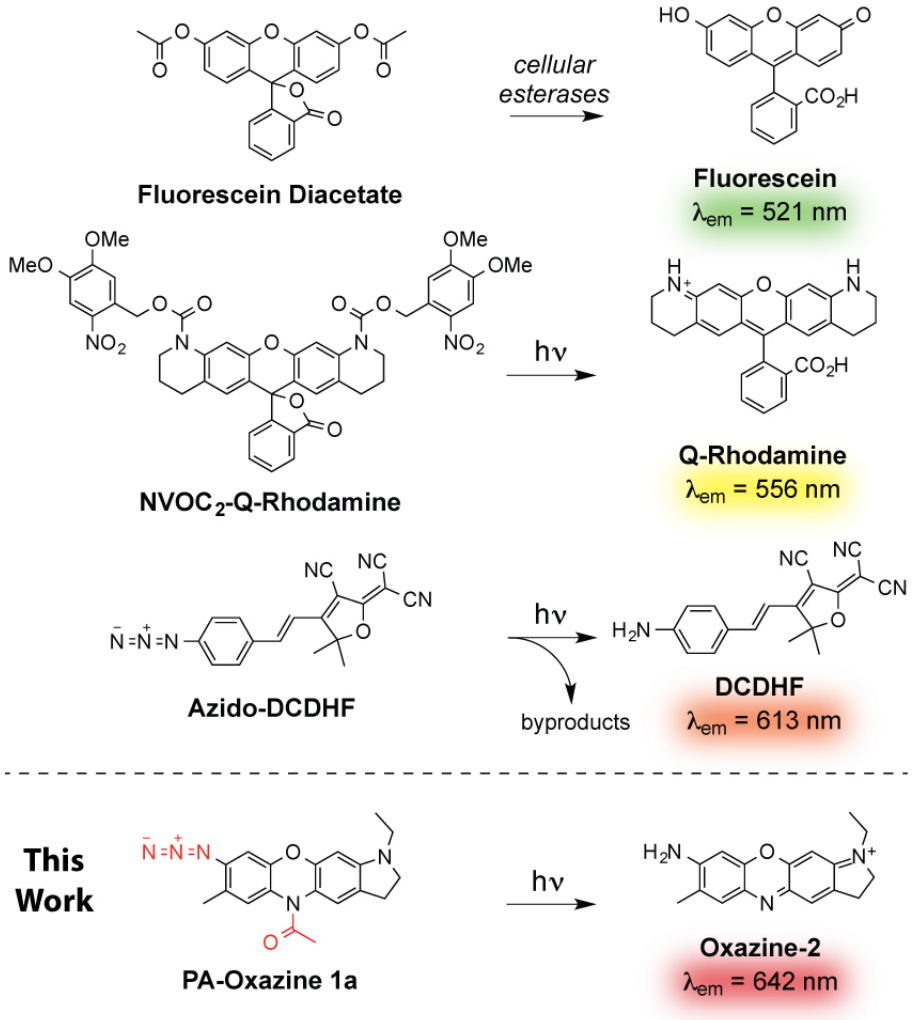


Figure 5-1. Strategies for caging fluorescent dyes. Approaches include trapping molecules in non-fluorescent states with enzymatically cleavable or photocleavable capping groups. Some pro-fluorescent molecules are intrinsically photoactivatable. The latter category most closely describes our approach to synthesize photoactivatable oxazine fluorophores.

Previously, other fluorophore scaffolds have been derivatized to overcome challenges of solubility[15], cell-permeability[16], and brightness[17], and to introduce functionality such as photoactivatability[18,19]. Representative examples include fluorescein diacetate, which is cell permeable and cleaved by intracellular esterases[20] to generate fluorescent product, and the photoactivatable fluorophores $NVOC_2$-Rhodamine[21] and azido-DCDHF[22] (**Fig. 5-1**). While NVOC is commonly used as a photocleavable group, it incorporates unwanted bulk and hydrophobicity into the probe. Also, in the case of $NVOC_2$-Rhodamine, full uncaging of the fluorophore requires two photocleavage events. Azido-DCDHF contains a smaller photoactivation motif that minimally perturbs its overall structure and requires just a single uncaging event; however, the photoreaction gives rise to a mixture of products.

Here, we report the design and synthesis of new oxazine derivatives and demonstrate their feasibility for use in live-cell imaging. First, we have addressed the synthetic chemistry challenges by employing recent advances in transition metal catalysis to synthesize acyl oxazine intermediates, providing an avenue for facile functionalization of the oxazine scaffold. Second, we prepared azido-acyl oxazines and demonstrated that they undergo rapid and clean photoconversion to the oxazine fluorophores through cleavage of the acyl moiety accompanied by the concomitant release of molecular nitrogen. Lastly, using the TMP chemical tagging technology, we show that azido-acyl oxazines are useful reagents for live mammalian cell imaging, especially when compared to the native uncaged oxazine fluorophores.

5.2 Results

5.2.1 Chemical synthesis of azido-acyl oxazines

In order to apply the approaches displayed in **Figure 5-1** to the oxazine class of fluorophores, we chose to build upon the report that acyl oxazines undergo UV-induced cleavage of the N-acyl bond to generate oxazine dyes[23]. However, because acyl oxazines uncage relatively slowly, the extent of UV exposure required for efficient photoactivation is incompatible with living cells. Moreover, competitive Photo-Fries rearrangement products are obtained in addition to the desired oxazines. Inspired by azido-DCDHF, we set out to synthesize azido-acyl oxazines (**Fig. 5-1, PA-oxazine 1a**) in an attempt to enhance photoactivation. We anticipated the release of molecular nitrogen would provide a strong thermodynamic driving force for irreversibility of acyl bond cleavage, favoring the forward reaction and resulting in fewer radical recombination byproducts.

Previously, acyl oxazines have been prepared from their parent oxazine fluorophores by in situ reduction of the oxazine to its *leuco* form in the presence of an acylating agent (**Fig. 5-1**, Prior Works). However, this procedure is limited in its compatibility with other functionality, and is circuitous from a synthetic standpoint. To circumvent these challenges, we sought to take advantage of our recently reported diaryl ether scaffold[24] to prepare acyl oxazines directly without passing through the oxazine dye. Given the electronically activated nature of the diaryl ether system, we reasoned that bromination followed by coupling with primary amides or carbamates would furnish acyl oxazine products (**Fig. 5-2**).

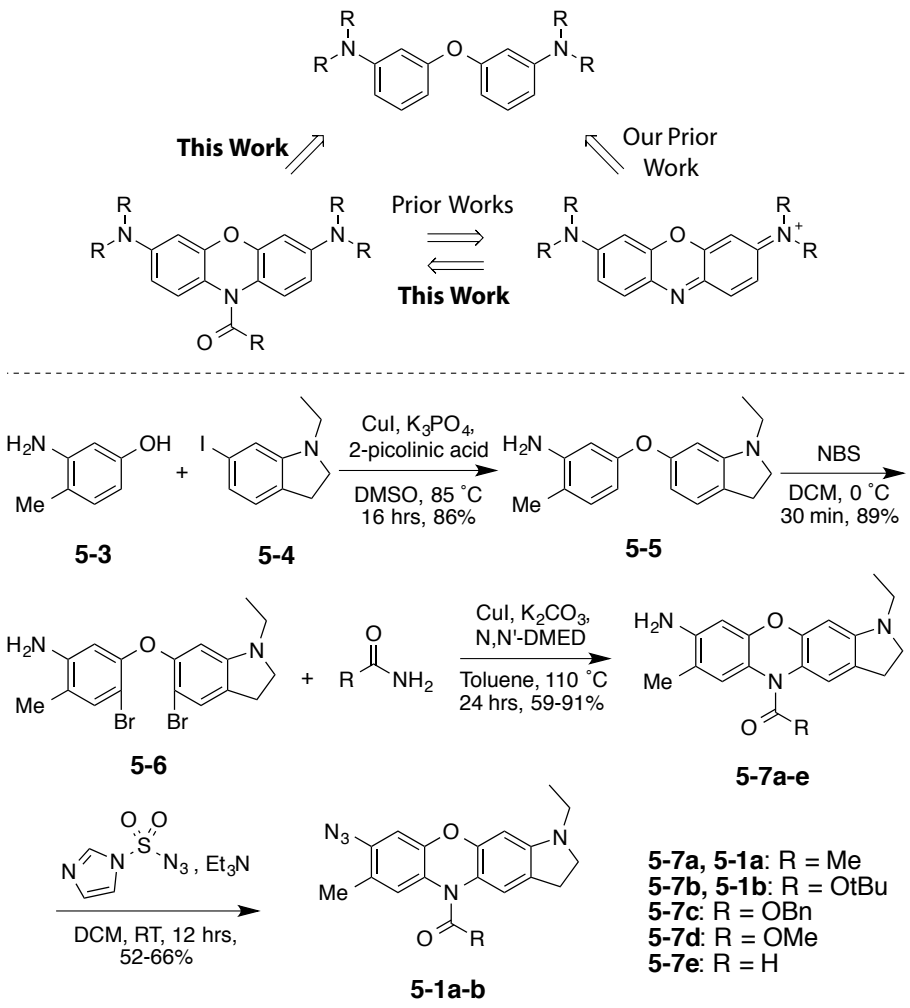


Figure 5-2. Synthesis of azido-acyl oxazines via acyl oxazine intermediates. N,N'-DMED = N,N'-dimethylelthylenedi-amine.

Diaryl ether **5-5** was prepared according to our previously reported strategy, utilizing a key copper(I) promoted reaction between aminophenol **5-3** and iodoindoline **5-4**. Bromination of **5-5** was achieved using N-bromosuccinimide, proceeding exclusively with the desired regioselectivity to afford **5-6** in high yield (**Fig. 5-2**). Then, a second key copper(I) promoted reaction[25] was implemented to construct the oxazine core, where the initial coupling between the amide or carbamate and **5-6** is followed by a second intramolecular coupling resulting in a cyclized acyl oxazine product. Importantly, this

reaction occurs in the presence of non-protected primary anilines, which are orthogonal under the conditions employed. Coupling/cyclization products bearing various acyl substituents were prepared in good to excellent yields, and several of these derivatives became colorful upon exposure to UV light. Of note, acyl oxazines can be deprotected under the appropriate conditions to yield leuco dyes, which upon exposure to air spontaneously oxidize to furnish the fluorescent oxazines. From acyl oxazines, synthesis of the target azido-acyl oxazines was achieved by diazotransfer from imidazole sulfonyl azide. Thus, this approach provides a general route to the construction of the oxazine core allowing for functionalization of the scaffold before deprotection of the fluorophore's sensitive conjugated cationic system.

5.2.2 Photochemical characterization of azido-acyl oxazines

As anticipated, azido-acyl oxazines were extremely sensitive to UV light and readily converted to fluorescent photoproducts after brief exposures (**Fig. 5-3a**). In the UV-Vis absorbance and fluorescence traces of a representative compound, **5-1b**, following increasing exposure to UV light, an absorbance band steadily arises in the visible region with a l_{max} of 612 nm, as does a fluorescence band with l_{max} centered at 642 nm. Closer inspection of the UV region within **Figure 5-3b** displays several isosbestic points for the photochemical process (**Fig. 5-4**), indicative of a clean reaction devoid of accumulating intermediates or byproducts.

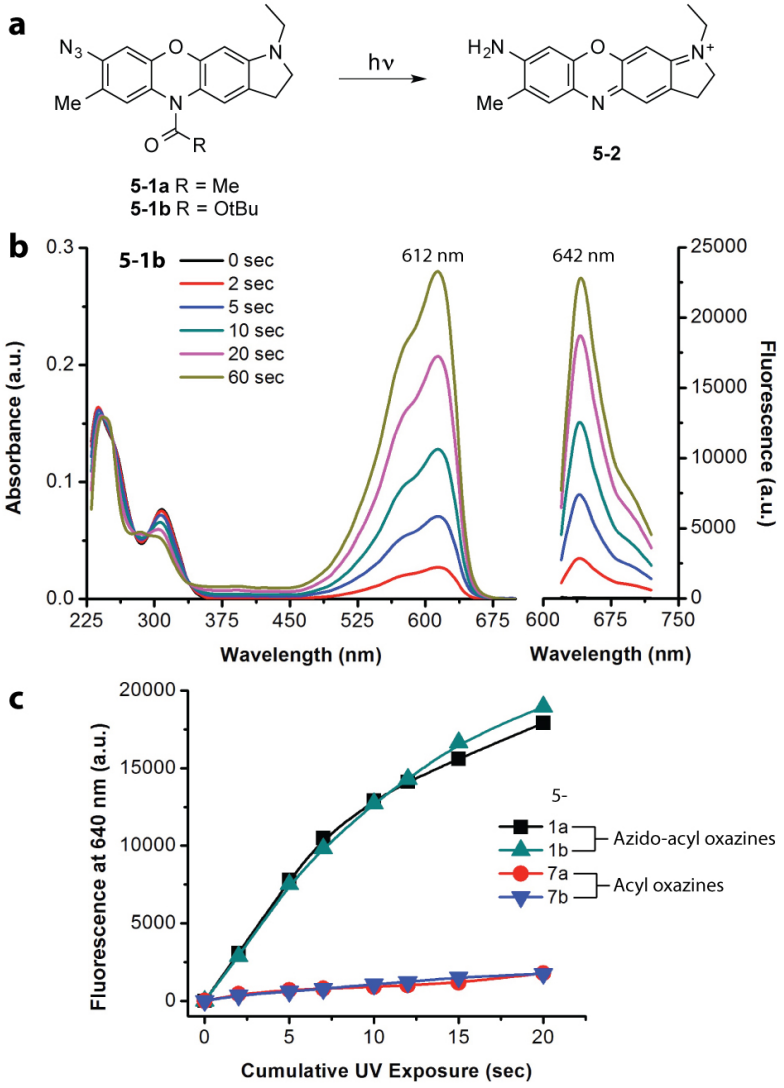


Figure 5-3. The azido-acyl oxazine uncaging photoreaction. (a) Photoactivation of azido-acyl oxazines and acyl oxazines. (b) Absorbance and fluorescence spectra (excitation = 594 nm) of **5-1b** as a function of total exposure to 254 nm light. (c) Comparison of photoactivation rates for the azido-acyl oxazines **5-1a** and **5-1b** vs. the corresponding acyl oxazines **5-7a** and **5-7b**.

The absorbance and fluorescence traces of the photoproducts overlay nearly identically with those of the independently synthesized oxazine **5-2**, which is the expected product of the photoreaction (**Fig. 5-5**). Additionally, HPLC analysis of the crude photoreaction indicated formation of just a single major product, identified as the anticipated oxazine by co-elution with the independently synthesized authentic marker **5-**

2 (**Fig. 5-6**). Together, these results demonstrate that the photoreaction proceeds cleanly to a single product, and that the product is the oxazine fluorophore **5-2** generated by cleavage of the acyl group and release of molecular nitrogen from the parent compound. In comparison to the acyl oxazines **5-7a** and **5-7b**, the corresponding azido-acyl oxazine derivatives **5-1a** and **5-1b** undergo more rapid uncaging to the fluorophore in response to UV light, with rate accelerations of 20-30 fold (**Fig. 5-3c**).

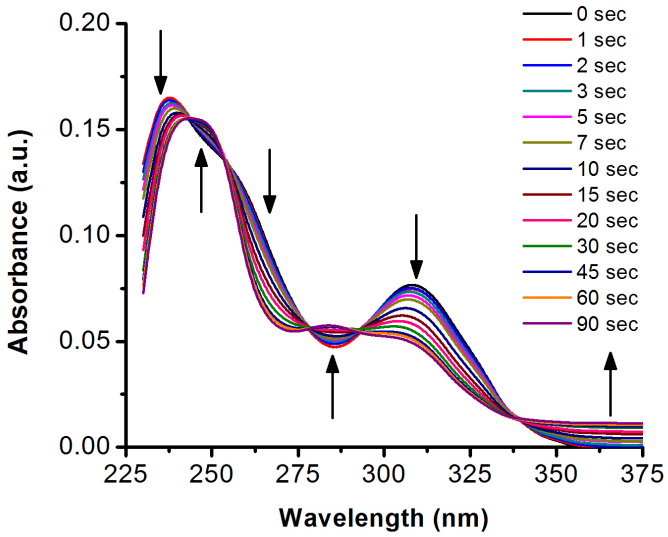


Figure 5-4. UV absorbance spectra of the photoreaction. The spectra were obtained for compound **5-1b** after varying durations of exposure to 254 nm light. Focus on the UV-region.

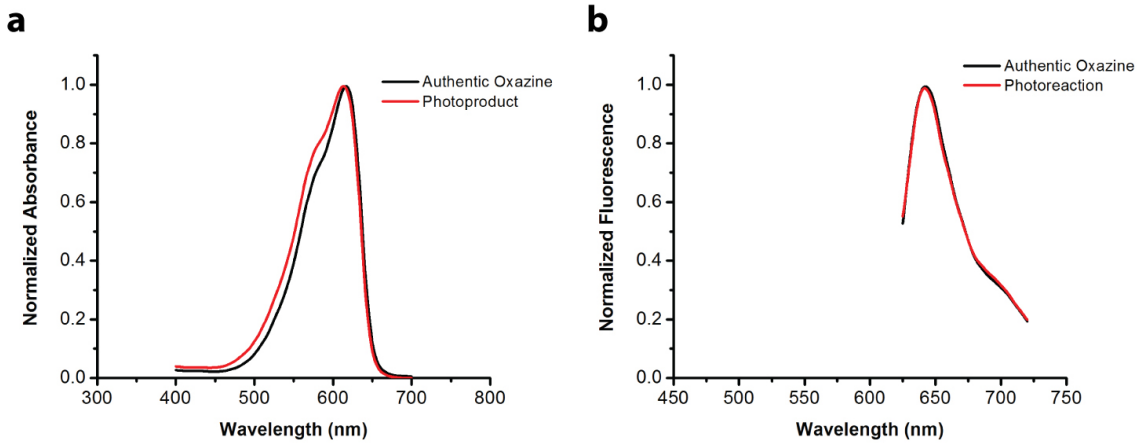


Figure 5-5. UV-vis spectra of the photoproduct. Overlay of (**a**) absorbance and (**a**) fluorescence spectra for the **5-1b** photoproduct and authentic oxazine **5-2**. Fluorescence excitation = 594 nm.

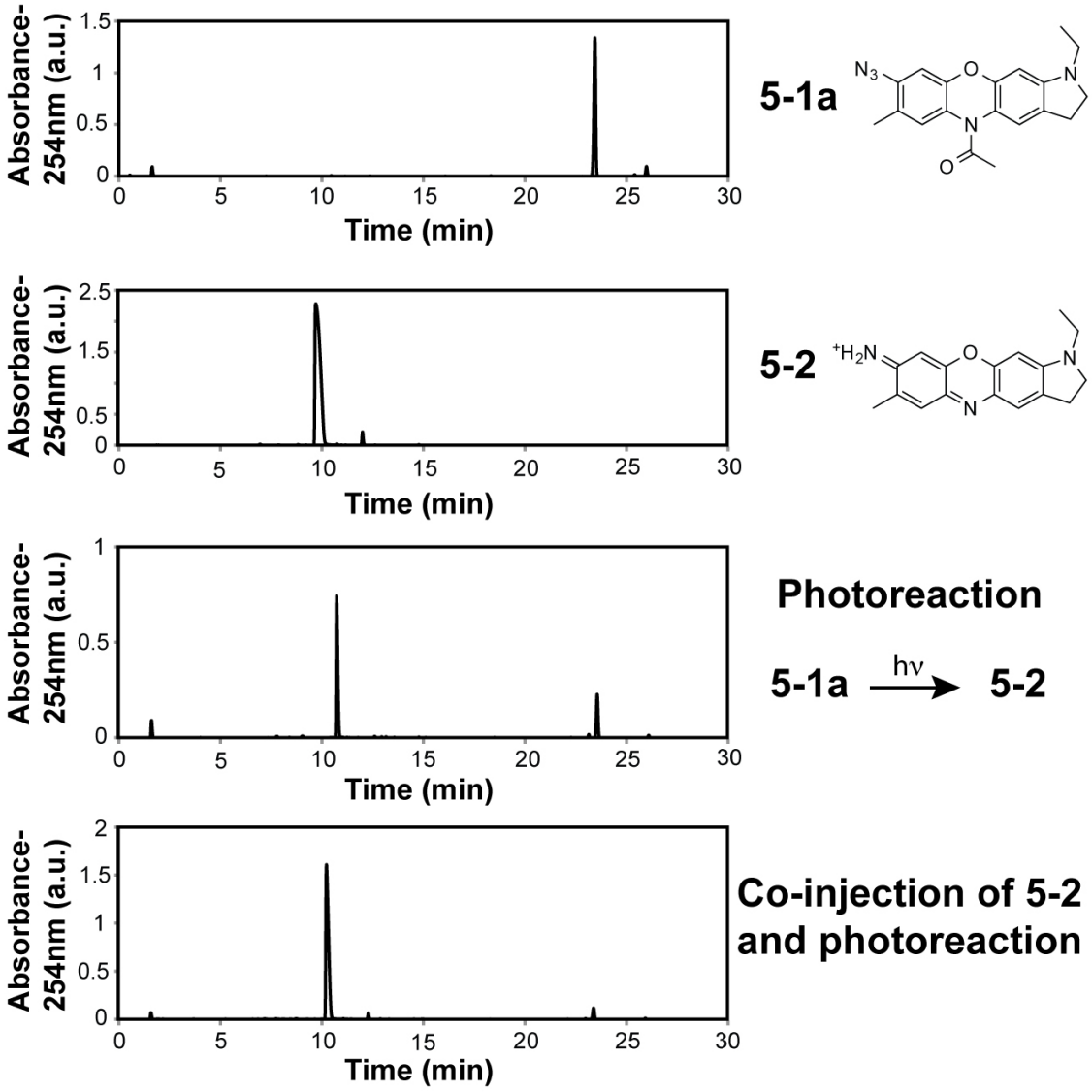


Figure 5-6. Reverse-phase HPLC analysis of the azido-acyl oxazine photoreaction. Co-injection of the authentic oxazine **5-2** and the photoreaction (lower panel) shows a single peak, indicating co-elution of the photoproduct with the oxazine fluorophore

5.2.3 Live-cell imaging with photoactivatable azido-acyl oxazines

To assess whether the photoactivatable oxazines were viable reagents for live cell imaging, we first tested the azido-acetyl oxazine derivative **5-1a** for permeability and photoactivatability in living cells. HeLa cells incubated with 2 μM of **5-1a** were irradiated with 365 nm light for 1s, resulting in a dramatic increase in the fluorescence signal (**Fig. 5-7**). This demonstrated that the photoactivatable oxazine is cell permeable and capable of undergoing photoactivation within the cellular environment. Given these promising results, we then tested the trimethoprim (TMP) conjugate of the photoactivatable oxazine for labeling E. coli dihydrofolate reductase (eDHFR) in HEK293T cells (**Fig. 5-8**). TMP binds to eDHFR with nanomolar affinity, allowing specific labeling of a protein of interest (POI) when it is fused to eDHFR by staining with TMP-fluorophore conjugates (**Fig. 5-8A**).

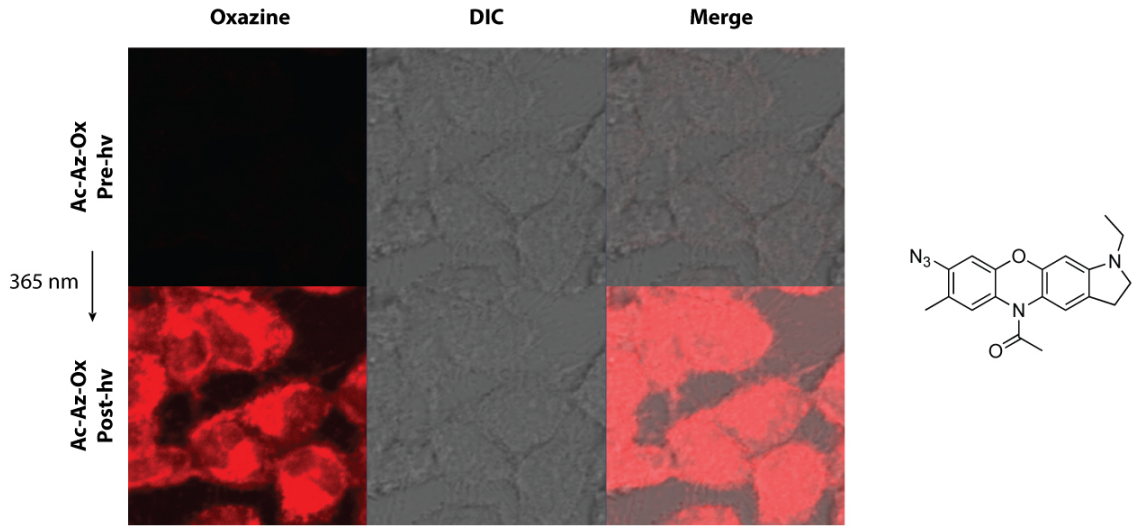


Figure 5-7. Photoactivation and fluorescence imaging of azido-acyl oxazines in live mammalian cells. The azido-acyl oxazine **5-1a** was photoactivated and imaged in living HeLa cells.

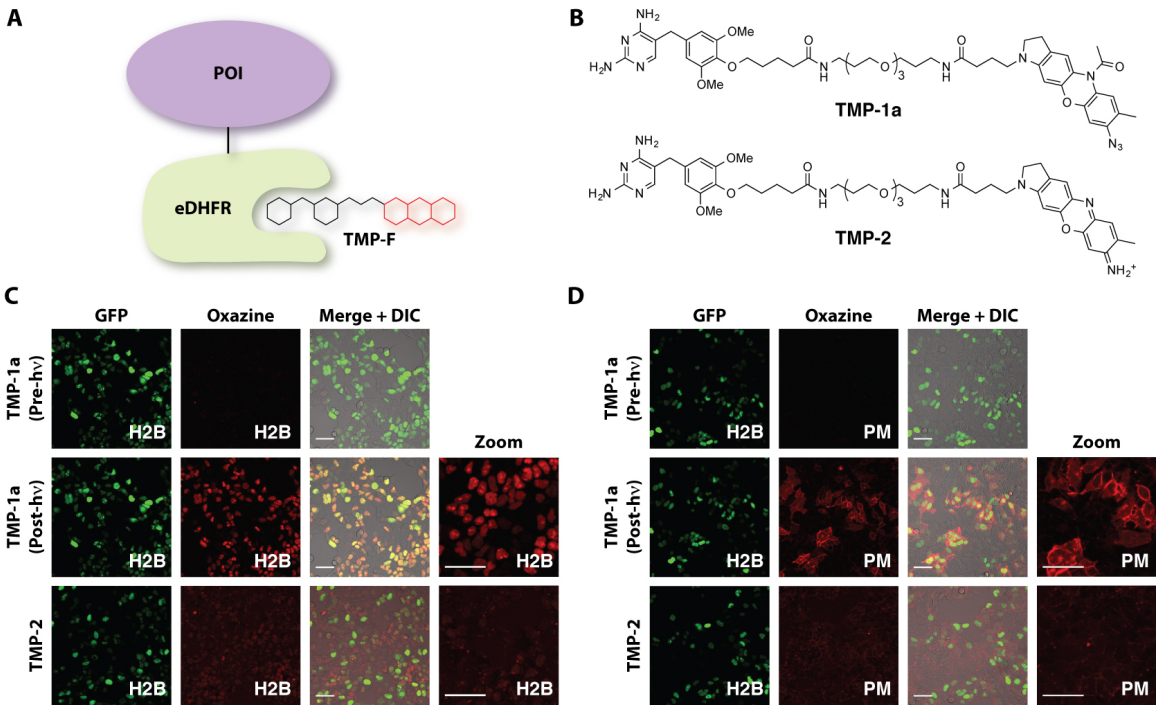


Figure 5-8. Protein-specific labeling in live cells with TMP-oxazine conjugates. (**A**) eDHFR is fused to a protein of interest (POI) and labeled with trimethoprim based probes. (**B**) Structures of the photoactivatable **TMP-1a** and the oxazine fluorophore **TMP-2**. (**C**) Labeling of a nuclear localized H2B-eDHFR fusion and (**D**) a plasma membrane (PM) localized eDHFR in HEK cells, co-tranfected with H2B-GFP. Scale bar = 50 μm. Photoactivation conducted with 365nm light.

We prepared the TMP conjugate of the photoactivatable acetyl-azido oxazine (**TMP-1a**) as well as the native oxazine fluorophore (**TMP-2**) as a point of comparison (**Fig. 5-8B**). Labeling of nuclear localized H2B-eDHFR with **TMP-1a** followed by a brief exposure to 365 nm light resulted in bright fluorescent labeling of the target protein (**Fig. 5-8C**). Interestingly, when **TMP-2** was imaged, very weak staining was observed, likely on account of poor cell permeability. We also observed clean labeling of a plasma membrane (PM) localized eDHFR with the photoactivatable **TMP-1a**, whereas **TMP-2** showed only weak staining of the intracellular DHFR (**Fig. 5-8D**). Of note, we have observed similarly poor cellular behavior with other oxazine analogs (data not shown).

These results indicate that the caged azido-acetyl oxazine is a superior reagent for live cell applications when compared to its corresponding oxazine fluorophore.

5.3 Discussion

In conclusion, we present a new synthetic route to the synthesis and derivatization of oxazine dyes and demonstrate the utility of this approach through the design and synthesis of a photoactivatable motif for oxazine fluorophores. We demonstrate that these reagents undergo rapid and clean photoconversion to a single major fluorescent product, and further apply these reagents for live cell imaging. Importantly, caged oxazines were found to be useful alternatives to native oxazine dyes on account of their improved cellular behavior. This new class of reagents expands our toolkit for probing the dynamics of individual molecules and unraveling the molecular mechanisms of biology in living cells.

5.4 Experimental Methods

General procedure 5.4.1 for the Copper(I) promoted coupling of phenols and aryl iodides.

According to the published protocol by Maiti and Buchwald[26], an oven dried, sealable glass vessel was charged with a magenetic stirbar, the phenol (2.40 mmol), potassium phosphate (4.00 mmol, 849 mg), copper(I) iodide (0.20 mmol, 38 mg), 2-picolinic acid (0.40 mmol, 49 mg), and the aryl iodide, if a solid (2.00 mmol). The vessel was then fitted with a rubber septum, evacuated under vacuum and backfilled with argon. This process was repeated 3 times. The vessel was then charged with DMSO (4.0 mL), or if the aryl iodide is a liquid, the vessel was charged with the aryl iodide as a solution in

DMSO. The rubber septum was removed and the reaction vessel was immediately sealed tightly with a Teflon screw cap. The reaction was then heated to 85 ˚C for 16-24 hours. After cooling to room temperature, the reaction was diluted with 10 mL of water and extracted with ethyl acetate (25 mL, 4x). The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a crude residue. Purification by flash chromatography (hexanes/ethyl acetate) afforded the diaryl ethers as colorless oils, which were stored at -20 ˚C under inert atmosphere.

General procedure 5.4.2 for the Copper(I) promoted coupling and cyclization of di-arylbromides with amides and carbamates.

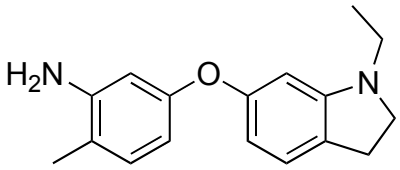
According to the published protocol by Klapars et al.[25], an oven dried, sealable glass vessel was charged with a magenetic stirbar, the di-bromoaryl ether (1.00 mmol), the amide or carbamate (1.20-1.50 mmol), potassium carbonate (3.00 mmol, 415 mg), copper(I) iodide (0.10 mmol, 19 mg), and N,N'-DMED (0.20 mmol, 17.6 mg). The vessel was then fitted with a rubber septum, evacuated under vacuum and backfilled with argon. This process was repeated 3 times. The vessel was then charged with toluene (1.2 mL). The rubber septum was removed and the reaction vessel was immediately sealed tightly with a Teflon screw cap. The reaction was then heated to 110 ˚C for 24 hours. After cooling to room temperature, the reaction was diluted with 10 mL of water and extracted with ethyl acetate (25 mL, 4x). The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a crude residue. Purification by flash chromatography (hexanes/ethyl acetate) afforded the cyclized acyl-leuco oxazines as colorless amorphous solids, which were stored protected from light at -20 ˚C under inert atmosphere.

General procedure 5.4.3 for converting acyl-oxazine to azido-acyl oxazines.

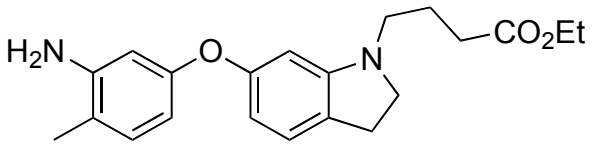
Imidazole sulfonyl azide · HCl (172 mg, 0.821 mmol) was dissolved in anhydrous dichloromethane (6 mL) and treated with triethylamine (0.32 mL, 2.30 mmol). To this mixture, the aniline (0.657 mmol) was added followed by copper (II) sulfate pentahydrate (~1 mg, 1.0 mol %). The reaction was stirred at room temperature for 6 hours and monitored for consumption of starting material. If the reaction did not reach completion, it was supplemented with additional triethylamine and imidazole sulfonyl azide · HCl. Upon completion, the reaction was filtered through celite, then diluted with 10 mL of water and extracted with dichloromethane (25 mL, 4x). The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a crude residue. Purification by flash chromatography (hexanes/ethyl acetate) afforded the azido-acyl oxazines as colorless amorphous solids, which were stored protected from light at -20 ˚C under inert atmosphere.

Cell culture, transfection and labeling. HEK293T cells were cultured in DMEM with glutamine (Bibco), supplemented with 10% v/v fetal bovine serum and 1% v/v Pen/Strep. Cells were maintained under 5% CO_2 at 37 ˚C. For transfection, labeling, and fluorescence imaging, cells were plated in 8-well chambered #1 borosilicate coverglass (ThermoFisher) and grown for 24 h before being transfected. Transfection of plasmids (0.4 ug DNA per well) expressing H2B-GFP and eDHFR fusion proteins of interest was achieved with Fugene HD (Roche). 24 h following transfection, 300 uL of fresh media containing 0.3 uL of the fluorophore stock solution (1 mM in DMF) was added to the well (1 uM final concentration of fluorophore). Cells were incubated with the staining solution for 10 min at 37 ˚C then washed with fresh media twice before imaging.

Confocal images were obtained on a Zeiss LSM700 confocal and processed with ZEN software. Photoactivation of the azido-acyl oxazines was achieved by illumination with 365 nm excitation.

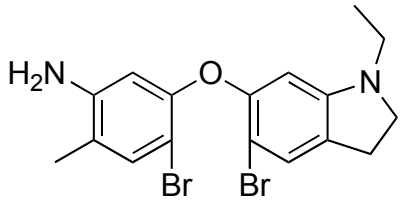


5-(1-ethylindolin-6-yloxy)-2-methylaniline (5-5). Following general procedure **5.4.1**, 3-amino-4-methyl phenol (1.13 g, 9.18 mmol) and 1-ethyl-6-iodoindoline **4-2c** (2.09 g, 7.65 mmol) were coupled to provide **5-5** (1.66 g, 81%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.00 (dd, *J* = 7.9, 1.4 Hz, 2H), 6.41 (dd, *J* = 8.1, 2.4 Hz, 1H), 6.38 (d, *J* = 2.4 Hz, 1H), 6.30 (dd, *J* = 7.8, 2.1 Hz, 1H), 6.22 (d, *J* = 2.1 Hz, 1H), 3.62 (br s, 2H), 3.41 (t, *J* = 8.2 Hz, 2H), 3.12 (q, *J* = 7.2 Hz, 2H), 2.96 (t, *J* = 8.2 Hz, 2H), 2.16 (s, 3H), 1.20 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 157.82, 157.57, 154.15, 145.96, 131.44, 125.42, 124.89, 117.02, 108.94, 107.99, 105.44, 99.58, 53.16, 43.28, 28.32, 17.06, 12.22. HRMS (FAB+) Calcd. For $C_{17}H_{20}N_2O^+$ [M^+]: 268.1576; found 268.1580.

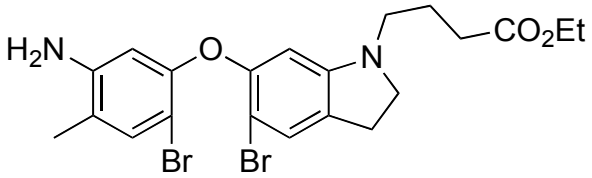


ethyl 4-(6-(3-amino-4-methylphenoxy)indolin-1-yl)butanoate (5-S2). Following general procedure **5.4.1**, 3-amino-4-methyl phenol (880 mg, 7.15 mmol) and **5-S1** (2.14 g, 5.96 mmol) were coupled to provide **5-S2** (1.69 g, 80%) as a colorless oil. ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.98 (dd, *J* = 7.8 Hz, 3.1 Hz, 2H), 6.39 (dd, *J* = 7.8, 2.3 Hz, 1H), 6.38 (s, 1H), 6.28 (dd, *J* = 7.8, 2.1 Hz, 1H), 6.19 (d, *J* = 2.1 Hz, 1H), 4.14 (q, *J* = 7.1 Hz, 2H), 3.57 (br s, 2H), 3.42 (t, *J* = 8.3 Hz, 2H), 3.08 (t, *J* = 7.1 Hz, 2H), 2.96 (t, *J* = 8.2 Hz, 2H), 2.41 (t, *J* = 7.4 Hz, 2H), 2.15 (s, 3H), 1.94 (p, *J* = 7.3 Hz, 2H), 1.27 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.74, 157.86,

157.46, 154.29, 145.90, 131.44, 125.00, 124.91, 117.10, 108.99, 107.96, 105.49, 99.27, 60.79, 53.99, 48.76, 32.20, 28.34, 23.22, 17.05, 14.61. HRMS (FAB+) Calcd. For $C_{21}H_{26}N_2O_3^+$ [M^+]: 354.1943; found 354.1946.

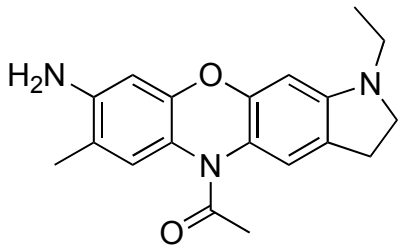


4-bromo-5-(5-bromo-1-ethylindolin-6-yloxy)-2-methylaniline (5-6). Compound **5-5** (1.58 g, 5.91 mmol) was dissolved in dichloromethane (20 mL) and cooled to 0˚C in an ice bath. The solution was then treated with N-bromosuccinimide (2.16 g, 12.1 mmol) in small portions over a period of 10 minutes. After stirring at 0˚C for 30 minutes, the reaction showed completion by TLC and was treated with 25 mL of sat. aq. $NaHCO_3$. The organic layer was separated from the aqueous, and the aqueous layer extracted twice with 25 mL of dichloromethane. The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a red-brown oil. Purified by flash chromatography to yield 1.91 g of **5-6** as a beige solid (89%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.26 (s, 1H), 7.23 (s, 1H), 6.14 (s, 1H), 6.08 (s, 1H), 3.59 (br s, 2H), 3.41 (t, *J* = 8.3 Hz, 2H), 3.05 (q, *J* = 7.2 Hz, 2H), 2.97 (t, *J* = 8.3 Hz, 2H), 2.13 (s, 3H), 1.14 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 153.39, 153.15, 152.91, 145.25, 134.71, 128.70, 128.15, 119.04, 105.20, 101.03, 100.91, 100.50, 52.84, 43.12, 28.11, 16.88, 12.00. HRMS (FAB+) Calcd. For $C_{17}H_{18}Br_2N_2O^+$ [M^+]: 423.9786; found 423.9781.

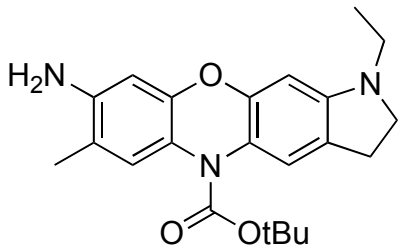


ethyl 4-(6-(5-amino-2-bromo-4-methylphenoxy)-5-bromoindolin-1-yl)butanoate (5-S3). Compound **5-S2** (1.58 g, 5.91 mmol) was dissolved in dichloromethane (20 mL) and cooled to 0˚C in an ice bath. The solution was then treated with N-bromosuccinimide (1.62 g, 9.08 mmol)

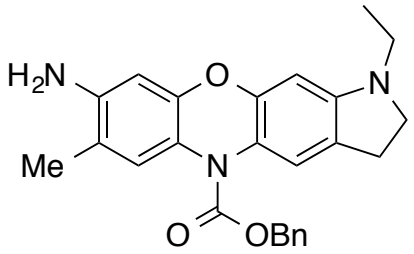
in small portions over a period of 10 minutes. After stirring at 0˚C for 30 minutes, the reaction showed completion by TLC and was treated with 25 mL of sat. aq. $NaHCO_3$. The organic layer was separated from the aqueous, and the aqueous layer extracted twice with 25 mL of dichloromethane. The combined organic layers were washed with brine and dried over Na_2SO_4, then concentrated in vacuo to a red-brown oil. Purified by flash chromatography to yield 1.98 g of **5-S3** as an amber oil (87%). HRMS (FAB+) Calcd. For $C_{21}H_{24}Br_2N_2O_3^+$ [M^+]: 510.0154; found 510.0170.



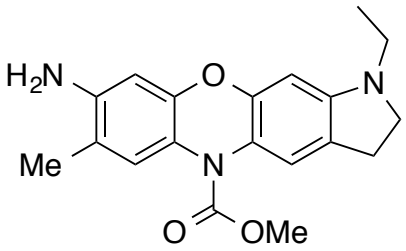
1-(8-amino-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazin-5(1*H*)-yl)ethan-1-one (5-7a). Following general procedure **5.4.2**, **5-6** (319 mg, 0.75 mmol) and acetamide (53 mg, 0.90 mmol) were coupled to provide **5-7a** (218 mg, 90%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.13 (s, 2H), 6.44 (s, 1H), 6.21 (s, 1H), 3.64 (br s, 2H), 3.38 (t, *J* = 8.2 Hz, 2H), 3.12 (q, *J* = 7.2 Hz, 2H), 2.95 (t, *J* = 8.2 Hz, 2H), 2.29 (s, 3H), 2.16 (s, 3H), 0.91 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (300 MHz, $CDCl_3$) δ 170.13, 151.69, 151.49, 150.86, 143.48, 126.63, 124.94, 121.44, 121.13, 119.88, 117.15, 103.06, 96.20, 53.01, 43.40, 28.38, 23.27, 17.37, 12.10. HRMS (FAB+) Calcd. For $C_{19}H_{21}N_3O_2^+$ [M^+]: 323.1634; found 323.1644.



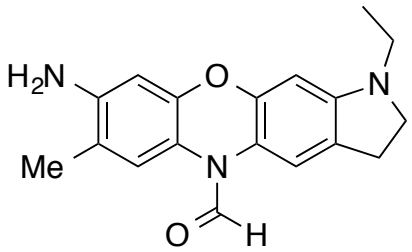
***tert*-butyl 8-amino-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazine-5(1*H*)-carboxylate (5-7b).** Following general procedure **5.4.2**, **5-6** (319 mg, 0.75 mmol) and t-butyl carbamate (106 mg, 0.90 mmol) were coupled to provide **5-7b** (204 mg, 71%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.20 (s, 2H), 6.37 (s, 1H), 6.16 (s, 1H), 3.57 (s, 2H), 3.35 (t, *J* = 7.8 Hz, 2H), 3.11 (q, *J* = 7.1 Hz, 2H), 2.93 (t, *J* = 7.6 Hz, 2H), 2.14 (s, 3H), 1.55 (s, 9H), 1.19 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 153.29, 150.82, 150.77, 149.98, 142.72, 126.79, 124.72, 121.23, 120.67, 119.25, 116.98, 102.83, 96.03, 81.71, 53.11, 43.58, 28.68, 28.44, 17.31, 12.21. HRMS (FAB+) Calcd. For $C_{22}H_{27}N_3O_3^+$ [M^+]: 381.2052; found 381.2058.



benzyl 8-amino-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazine-5(1*H*)-carboxylate (5-7c). Following general procedure **5.4.2**, **5-6** (319 mg, 0.75 mmol) and benzyl carbamate (136 mg, 0.90 mmol) were coupled to provide **5-7c** (199 mg, 64%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.47 – 7.33 (m, 5H), 7.25 (d, *J* = 4.3 Hz, 2H), 6.39 (s, 1H), 6.20 (s, 1H), 5.30 (s, 2H), 3.60 (br s, 2H), 3.37 (t, *J* = 8.2 Hz, 2H), 3.12 (q, *J* = 7.2 Hz, 2H), 2.94 (t, *J* = 8.2 Hz, 2H), 2.13 (s, 3H), 1.21 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 154.31, 151.17, 150.91, 150.06, 143.24, 136.70, 128.90, 128.52, 128.43, 128.22, 126.62, 124.94, 120.12, 118.80, 117.17, 102.89, 96.05, 68.17, 53.06, 43.50, 28.42, 17.30, 12.20. HRMS (FAB+) Calcd. For $C_{25}H_{25}N_3O_3^+$ [M^+]: 415.1896; found 415.1907.

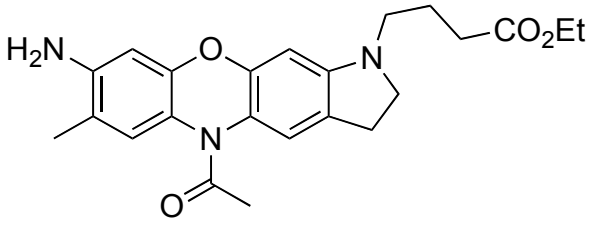


methyl 8-amino-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazine-5(1*H*)-carboxylate (5-7d). Following general procedure **5.4.2**, **5-6** (319 mg, 0.75 mmol) and methyl carbamate (68 mg, 0.90 mmol) were coupled to provide **5-7d** (168 mg, 66%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.20 (d, *J* = 5.7 Hz, 2H), 6.39 (s, 1H), 6.17 (s, 1H), 3.83 (s, 3H), 3.60 (br s, 2H), 3.37 (t, *J* = 8.1 Hz, 2H), 3.11 (q, *J* = 7.2 Hz, 2H), 2.95 (t, *J* = 8.1 Hz, 2H), 2.15 (s, 3H), 1.19 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 155.07, 151.16, 150.92, 150.11, 143.14, 126.48, 124.95, 120.97, 120.30, 118.82, 117.22, 102.88, 96.03, 53.54, 53.05, 43.47, 28.43, 17.37, 12.13. HRMS (FAB+) Calcd. For $C_{19}H_{21}N_3O_3^+$ [M^+]: 339.1583; found 339.1578.

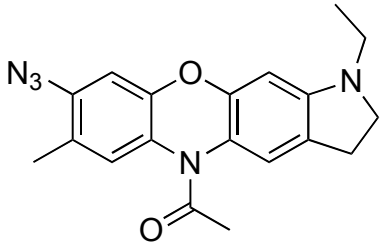


8-amino-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazine-5(1*H*)-carbaldehyde (5-7e). Following general procedure **5.4.2**, **5-6** (319 mg, 0.75 mmol) and formamide (51 mg, 1.13 mmol) were coupled to provide **5-7e** (137 mg, 59%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.52 (s, 1H), 7.64 (d, *J* = 5.3 Hz, 1H), 6.80 (d, *J* = 6.5 Hz, 1H), 6.38 (d, *J* = 5.7 Hz, 1H), 6.16 (d, *J* = 6.4 Hz, 1H), 3.54 (br s, 2H), 3.36 (t, *J* = 7.8 Hz, 2H), 3.11 (q, *J* = 7.2 Hz, 2H), 2.93 (t, *J* = 7.8 Hz, 2H), 2.13 (s, 3H), 1.17 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 160.05, 151.49, 148.70, 147.81, 143.74, 125.75, 124.78, 120.01, 119.26, 117.44, 114.78, 102.78,

96.00, 52.90, 43.33, 28.38, 17.30, 12.02. HRMS (FAB+) Calcd. For $C_{18}H_{19}N_3O_2^+$ [M^+]: 309.1477; found 309.1473.

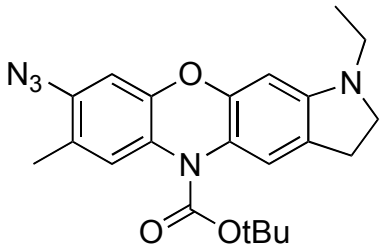


ethyl 4-(5-acetyl-8-amino-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazin-1(5*H*)-yl)butanoate (5-S4). Following general procedure **5.4.2**, **5-S3** (512 mg, 1.00 mmol) and acetamide (71 mg, 1.20 mmol) were coupled to provide **5-S4** (310 mg, 76%) as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.08 (m, 2H), 6.39 (s, 1H), 6.17 (s, 1H), 4.13 (q, *J* = 7.1 Hz, 2H), 3.69 (s, 2H), 3.35 (t, *J* = 8.3 Hz, 2H), 3.05 (t, *J* = 7.0 Hz, 2H), 2.92 (t, *J* = 8.2 Hz, 2H), 2.39 (t, *J* = 7.3 Hz, 2H), 2.25 (s, 3H), 2.11 (s, 3H), 1.91 (p, *J* = 7.2 Hz, 2H), 1.25 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.68, 170.10, 151.68, 151.65, 150.78, 143.72, 126.51, 124.48, 121.19, 121.16, 119.88, 117.09, 102.97, 95.89, 60.79, 53.88, 48.81, 32.06, 28.41, 23.24, 23.09, 17.35, 14.62. HRMS (FAB+) Calcd. For $C_{23}H_{27}N_3O_4^+$ [M^+]: 409.2002; found 409.2000.

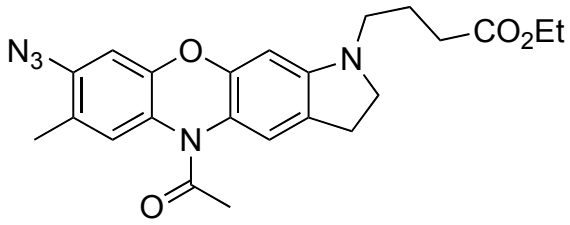


1-(8-azido-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazin-5(1*H*)-yl)ethan-1-one (5-1a). Following general procedure **5.4.3**, **5-7a** (167 mg, 0.516 mmol) was converted to **5-1a** (94

mg, 52%) and recovered as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (s, 1H), 7.06 (s, 1H), 6.88 (s, 1H), 6.22 (s, 1H), 3.41 (t, 2H), 3.15 (q, *J* = 7.2 Hz, 2H), 2.97 (t, *J* = 8.2 Hz, 2H), 2.30 (s, 3H), 2.20 (s, 3H), 1.21 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 170.00, 151.77, 151.48, 150.55, 136.63, 127.52, 127.01, 125.50, 124.47, 120.97, 119.11, 106.52, 96.08, 52.91, 43.26, 28.33, 23.29, 17.21, 12.07. HRMS (FAB+) Calcd. For $C_{19}H_{19}N_5O_2^+$ [M^+]: 349.1539; found 349.1539.

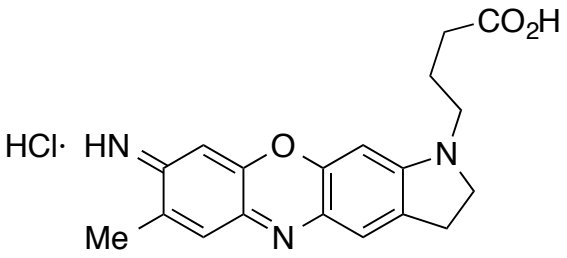


***tert*-butyl 8-azido-1-ethyl-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazine-5(1*H*)-carboxylate (5-1b).** Following general procedure **5.4.3**, **5-7b** (208 mg, 0.545 mmol) was converted to **5-1b** (146 mg, 66%) and recovered as a colorless amorphous solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.32 (s, 1H), 7.19 (s, 1H), 6.80 (s, 1H), 6.16 (s, 1H), 3.38 (t, *J* = 8.2 Hz, 2H), 3.13 (q, *J* = 7.2 Hz, 2H), 2.94 (t, *J* = 8.2 Hz, 2H), 2.18 (s, 3H), 1.54 (s, 9H), 1.20 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.72, 152.82, 151.23, 150.44, 135.74, 127.47, 126.35, 124.79, 124.06, 121.22, 118.47, 106.21, 95.59, 82.37, 53.91, 43.87, 28.60, 28.45, 17.15, 12.64. HRMS (FAB+) Calcd. For $C_{22}H_{25}N_5O_3^+$ [M^+]: 407.1957; found 407.1963.

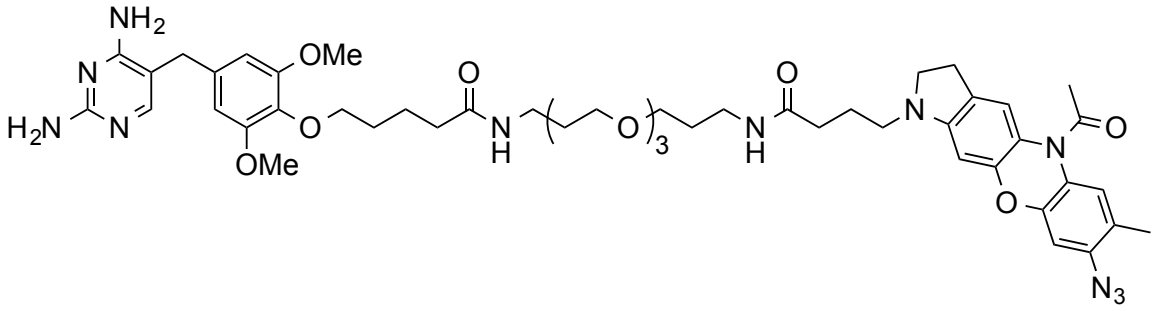


ethyl 4-(5-acetyl-8-azido-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazin-1(5*H*)-yl)butanoate (5-S5). Following general procedure **5.4.3**, **5-S4** (305 mg, 0.745 mmol) was converted to **5-S5** (184 mg, 57%) and recovered as a light-sensitive colorless amorphous solid. ^{1}H NMR (400 MHz,

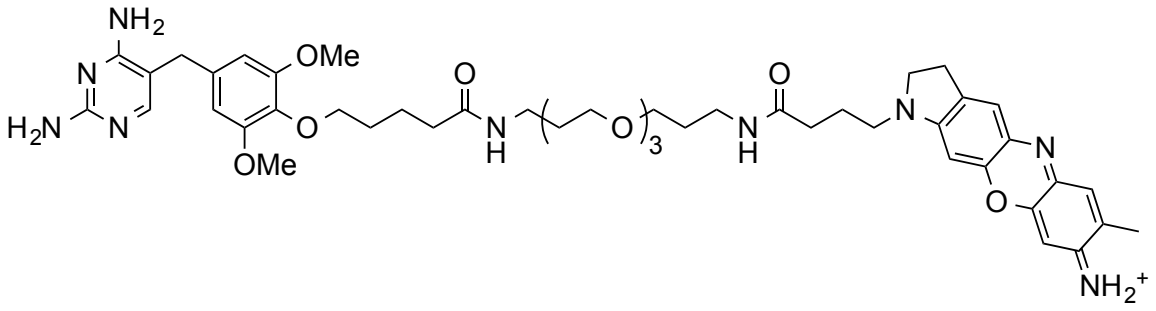
$CDCl_3$) δ 7.29 (s, 1H), 7.04 (s, 1H), 6.85 (s, 1H), 6.20 (s, 1H), 4.16 (q, *J* = 7.1 Hz, 2H), 3.42 (t, *J* = 8.3 Hz, 2H), 3.10 (t, *J* = 7.1 Hz, 2H), 2.97 (t, *J* = 8.3 Hz, 2H), 2.42 (t, *J* = 7.2 Hz, 2H), 2.29 (s, 3H), 2.19 (s, 3H), 1.94 (p, *J* = 7.2 Hz, 2H), 1.27 (t, *J* = 7.1 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 173.66, 169.98, 151.98, 151.47, 150.50, 136.61, 127.49, 126.94, 125.10, 124.44, 121.00, 119.09, 106.50, 95.84, 60.86, 53.80, 48.65, 32.01, 28.39, 23.30, 22.99, 17.25, 14.64. HRMS (FAB+) Calcd. For $C_{23}H_{25}N_5O_4^+$ [M^+]: 435.1907; found 435.1896.



4-(8-imino-7-methyl-2,3-dihydropyrrolo[3,2-*b*]phenoxazin-1(8*H*)-yl)butanoic acid · HCl salt (5-S6). Compound **5-S5** (25 g, 0.0574 mmol) was heated to 60 ˚C for 16 hours in a mixture of THF (3 mL) and 0.5N HCl (12 mL). The aqueous layer was extracted with 25 mL of a 4:1 mixture of ethyl acetate/hexanes, and then extracted with dichloromethane (3x 25 mL). The dichloromethane layers were concentrated under reduced pressure to yield **5-S6** as a deep blue residue (15 mg, 71%). ^{1}H NMR (400 MHz, MeOD) δ 7.49 (s, 1H), 7.37 (s, 1H), 6.76 (s, 1H), 6.69 (s, 1H), 4.11 (t, *J* = 6.0 Hz, 2H), 3.72 (t, *J* = 6.0 Hz, 2H), 3.31 (m, 2H), 2.49 (t, *J* = 6.8 Hz, 2H), 2.30 (s, 3H), 2.07 (p, *J* = 7.2 Hz, 2H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 175.27, 161.27, 158.18, 151.52, 147.62, 139.25, 136.91, 132.64, 132.05, 127.20, 126.70, 96.93, 91.14, 54.00, 46.86, 30.47, 25.99, 22.32, 16.36.

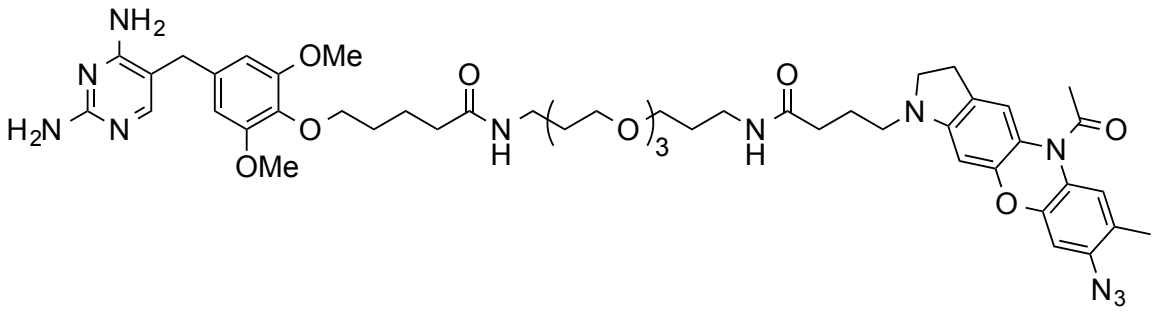


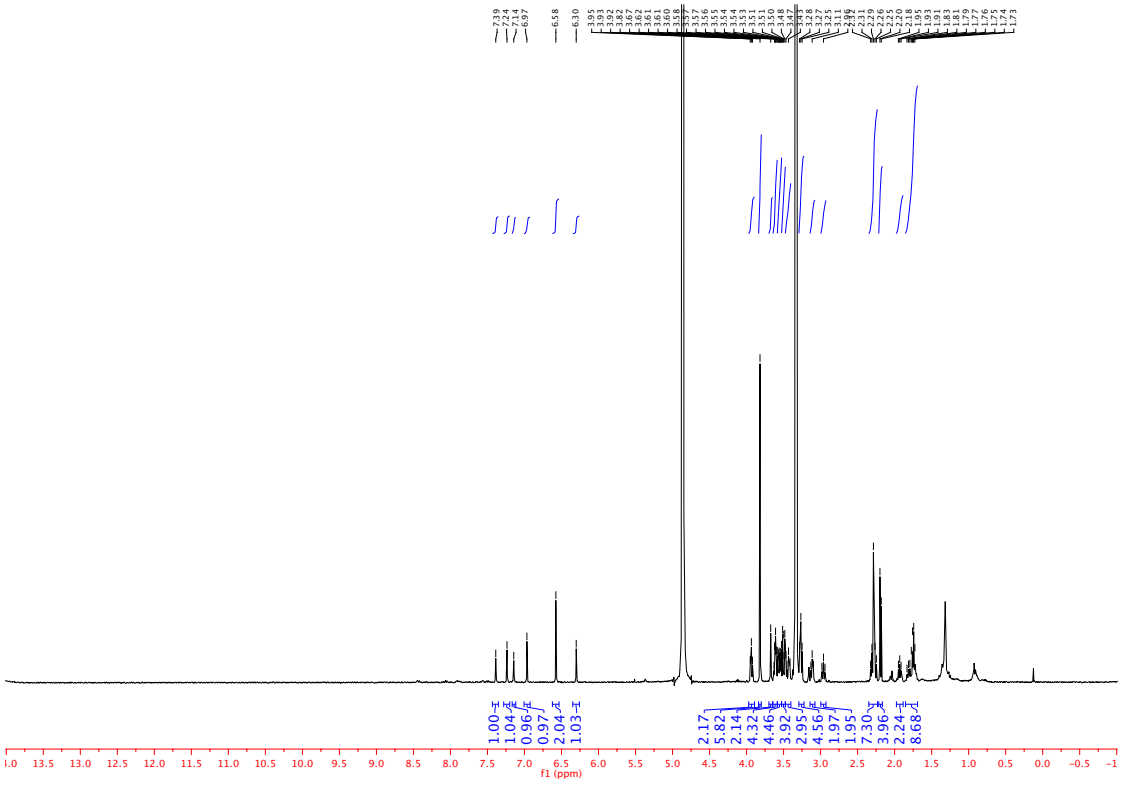
TMP-1a. Compound **5-S5** (25 mg, 0.0574 mmol) was subjected to hydrolysis conditions as above (THF/0.5N HCl) for 5 hours at 60 ˚C, then purified by flash chromatography (DCM/MeOH) to yield 8 mg of the carboxylic acid (35%). Then, TMP-$(PEG)_3$-NH_2 (1.2 mg, 1.7 umol), the carboxylic acid of **5-S5** (1.0 mg, 2.3 umol), HCTU (4.5 mg, 11 umol), triethylamine (1.0 uL, 7.2 umol) and dimethylformamide (0.2 mL) were combined in an amber vial, which was then purged with argon and sealed. Stirred at room temperature for 75 minutes, then purified directly by reverse phase HPLC to yield 0.8 mg of **TMP-1a** (48%). ^{1}H NMR (400 MHz, MeOD) δ 7.39 (s, 1H), 7.24 (s, 1H), 7.14 (s, 1H), 6.97 (s, 1H), 6.58 (s, 2H), 6.30 (s, 1H), 3.93 (t, *J* = 6.1 Hz, 2H), 3.82 (s, 6H), 3.67 (s, 2H), 3.63 - 3.59 (m, 4H), 3.59-3.52 (m, 4H), 3.52 - 3.46 (m, 4H), 3.43 (t, *J* = 8.4 Hz, 2H), 3.27 (t, *J* = 6.8 Hz, 4H), 3.11 (t, *J* = 6.9 Hz, 2H), 2.96 (t, *J* = 8.3 Hz, 2H), 2.29 (s, 3H), 2.29 (p, *J* = 8.1 Hz, 4H), 2.20 (s, 3H), 1.94 (p, *J* = 7.6 Hz, 2H), 1.86 – 1.69 (m, 8H). HRMS (FAB+) Calcd. For $C_{49}H_{66}N_{11}O_{10}^+$ [M^+]: 968.4989; found 968.5004.

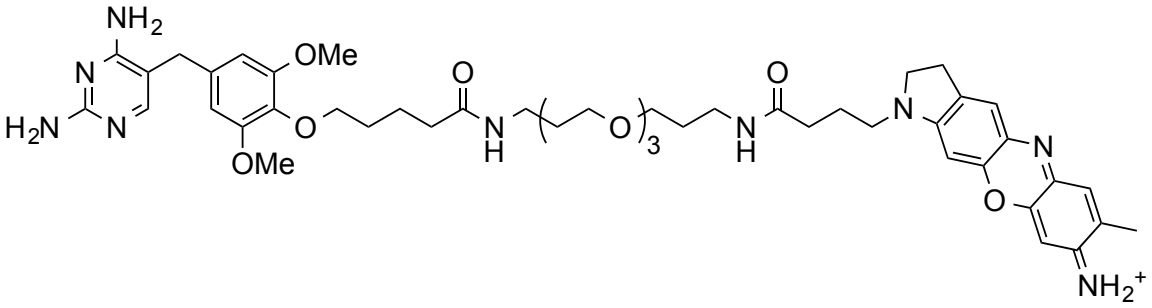


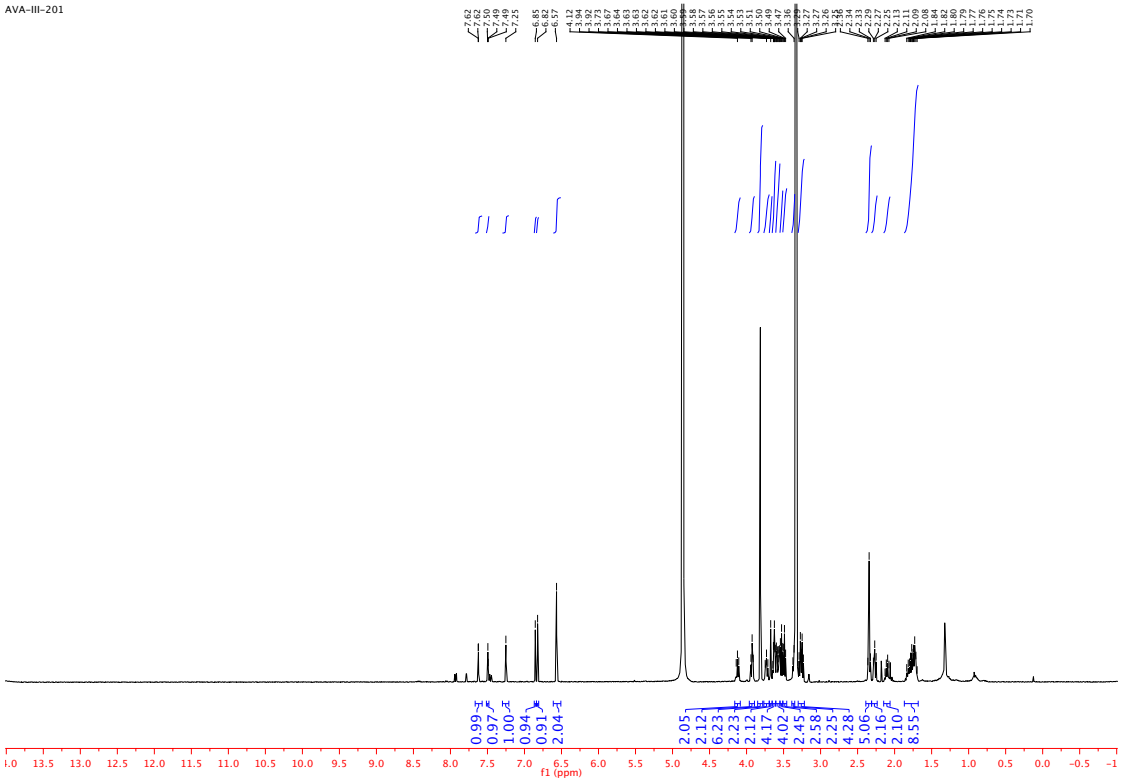
TMP-2. TMP-$(PEG)_3$-NH_2 (1.2 mg, 1.7 umol), **S09** (1.0 mg, 2.2 umol), HCTU (4.5 mg, 11 umol), triethylamine (1.0 uL, 7.2 umol) and dimethylformamide were combined in an amber vial, which was then purged with argon and sealed. Stirred at room temperature for 75 minutes, then purified directly by reverse phase HPLC to yield 1.2 mg of **TMP-2** (69%). ^{1}H NMR (400 MHz, MeOD) δ 7.62 (s, 1H), 7.49 (s, 1H), 7.25 (s, 1H), 6.85 (s, 1H), 6.82 (s, 1H), 6.57 (s, 2H), 4.12 (t, *J* = 6.9 Hz, 2H), 3.93 (t, *J* = 6.1 Hz, 2H), 3.81 (s, 6H), 3.73 (t, *J* = 7.1 Hz, 2H), 3.67 (s, 2H), 3.65 - 3.60 (m, 4H), 3.60 - 3.55 (m, 4H), 3.53 - 3.47 (m, 4H), 3.36 (m, 2H), 3.26 (dt, *J* = 9.4, 6.9 Hz, 4H), 2.34 (s, 3H), 2.34 (t, *J* = 6.8 Hz, 2H). 2.27 (t, *J* = 7.3 Hz, 2H), 2.10 (p, *J* = 7.2 Hz, 2H), 1.87 - 1.67 (m, 8H). HRMS (FAB+) Calcd. For $C_{47}H_{64}N_9O_9^+$ [M^+]: 898.4822; found 898.4857.

5.5 Spectra






NH2
N
H2N
N
OMe
O
OMe
O
N
H
O
3
N
H
O
N
N
O
NH2+


AVA-III-201
f1 (ppm)

5.6 References

1. Fernández-Suárez, M. & Ting, A. Y. Fluorescent probes for super-resolution imaging in living cells. *Nat. Rev. Mol. Cell Biol.* **9,** 929–943 (2008).
2. Day, R. N. & Davidson, M. W. The fluorescent protein palette: tools for cellular imaging. *Chem. Soc. Rev.* **38,** 2887–2921 (2009).
3. Jing, C. & Cornish, V. W. Chemical Tags for Labeling Proteins Inside Living Cells. *Acc. Chem. Res.* **44,** 784–792 (2011).
4. Chen, Z., Cornish, V. W. & Min, W. Chemical tags: inspiration for advanced imaging techniques. *Curr. Opin. Chem. Biol.* **17,** 637–643 (2013).
5. Miller, L. W., Cai, Y., Sheetz, M. P. & Cornish, V. W. In vivo protein labeling with trimethoprim conjugates: a flexible chemical tag. *Nat. Methods* **2,** 255–257 (2005).
6. Keppler, A. *et al.* A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nat. Biotechnol.* **21,** 86–89 (2003).
7. Los, G. V. & Wood, K. The HaloTag: a novel technology for cell imaging and protein analysis. *Methods Mol. Biol. Clifton NJ* **356,** 195–208 (2007).
8. Miyawaki, A., Griesbeck, O., Heim, R. & Tsien, R. Y. Dynamic and quantitative Ca2+ measurements using improved cameleons. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 2135–2140 (1999).
9. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3,** 793–795 (2006).
10. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).

11. Jing, C. & Cornish, V. W. A Fluorogenic TMP-Tag for High Signal-to-Background Intracellular Live Cell Imaging. *ACS Chem. Biol.* **8,** 1704–1712 (2013).

12. Vaughan, J. C., Jia, S. & Zhuang, X. Ultrabright photoactivatable fluorophores created by reductive caging. *Nat. Methods* **9,** 1181–1184 (2012).

13. Vogelsang, J., Cordes, T., Forthmann, C., Steinhauer, C. & Tinnefeld, P. Controlling the fluorescence of ordinary oxazine dyes for single-molecule switching and superresolution microscopy. *Proc. Natl. Acad. Sci.* **106,** 8107–8112 (2009).

14. Wombacher, R. *et al.* Live-cell super-resolution imaging with trimethoprim conjugates. *Nat. Methods* **7,** 717–719 (2010).

15. Romieu, A. *et al.* Postsynthetic derivatization of fluorophores with alpha-sulfo-beta-alanine dipeptide linker. Application to the preparation of water-soluble cyanine and rhodamine dyes. *Bioconjug. Chem.* **19,** 279–289 (2008).

16. Pan, D. *et al.* A general strategy for developing cell-permeable photo-modulatable organic fluorescent probes for live-cell super-resolution imaging. *Nat. Commun.* **5,** 5573 (2014).

17. Grimm, J. B. *et al.* A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nat. Methods* **12,** 244–250 (2015).

18. Kobayashi, T. *et al.* Highly Activatable and Rapidly Releasable Caged Fluorescein Derivatives. *J. Am. Chem. Soc.* **129,** 6696–6697 (2007).

19. Grimm, J. B. & Lavis, L. D. Synthesis of Rhodamines from Fluoresceins Using Pd-Catalyzed C–N Cross-Coupling. *Org. Lett.* **13,** 6354–6357 (2011).

20. Aeschbacher, M., Reinhardt, C. A. & Zbinden, G. A rapid cell membrane permeability test using fluorescent dyes and flow cytometry. *Cell Biol. Toxicol.* **2,** 247–255 (1986).

21. Wysocki, L. M. *et al.* Facile and general synthesis of photoactivatable xanthene dyes. *Angew. Chem. Int. Ed Engl.* **50,** 11206–11209 (2011).

22. Lord, S. J. *et al.* A photoactivatable push-pull fluorophore for single-molecule imaging in live cells. *J. Am. Chem. Soc.* **130,** 9204–9205 (2008).

23. Akiba, M., Dvornikov, A. S. & Rentzepis, P. M. Formation of oxazine dye by photochemical reaction of N-acyl oxazine derivatives. *J. Photochem. Photobiol. Chem.* **190,** 69–76 (2007).

24. Anzalone, A. V., Wang, T. Y., Chen, Z. & Cornish, V. W. A common diaryl ether intermediate for the gram-scale synthesis of oxazine and xanthene fluorophores. *Angew. Chem. Int. Ed Engl.* **52,** 650–654 (2013).

25. Klapars, A., Huang, X. & Buchwald, S. L. A General and Efficient Copper Catalyst for the Amidation of Aryl Halides. *J. Am. Chem. Soc.* **124,** 7421–7428 (2002).

26. Maiti, D. & Buchwald, S. L. Orthogonal Cu- and Pd-Based Catalyst Systems for the O- and N-Arylation of Aminophenols. *J. Am. Chem. Soc.* **131,** 17423–17429 (2009).

Appendix

A1. Identification and characterization of -1 PRF motifs from NGS data.

A pipeline was established for processing and analyzing NGS data described in **Chapter 3.2.1** from the point of generation to final motif identification Step 1 of the pipeline processes the raw sequencing fastq file and generates a file with every unique sequence and its read count. The HiSeq run yielded roughly 56 million sequencing reads after processing, from which >3 million unique sequences were recovered. Abundances (or read counts) of these sequences ranged from 1 read to >800,000 reads. 50% of the total sequencing reads are accounted for by the top 6,752 most abundant unique sequences (**Fig. A.1-1**). We set out to identify -1 PRF motifs from this dataset with a specific focus on RNA pseudoknots (PKs), which are known to promote -1 PRF.

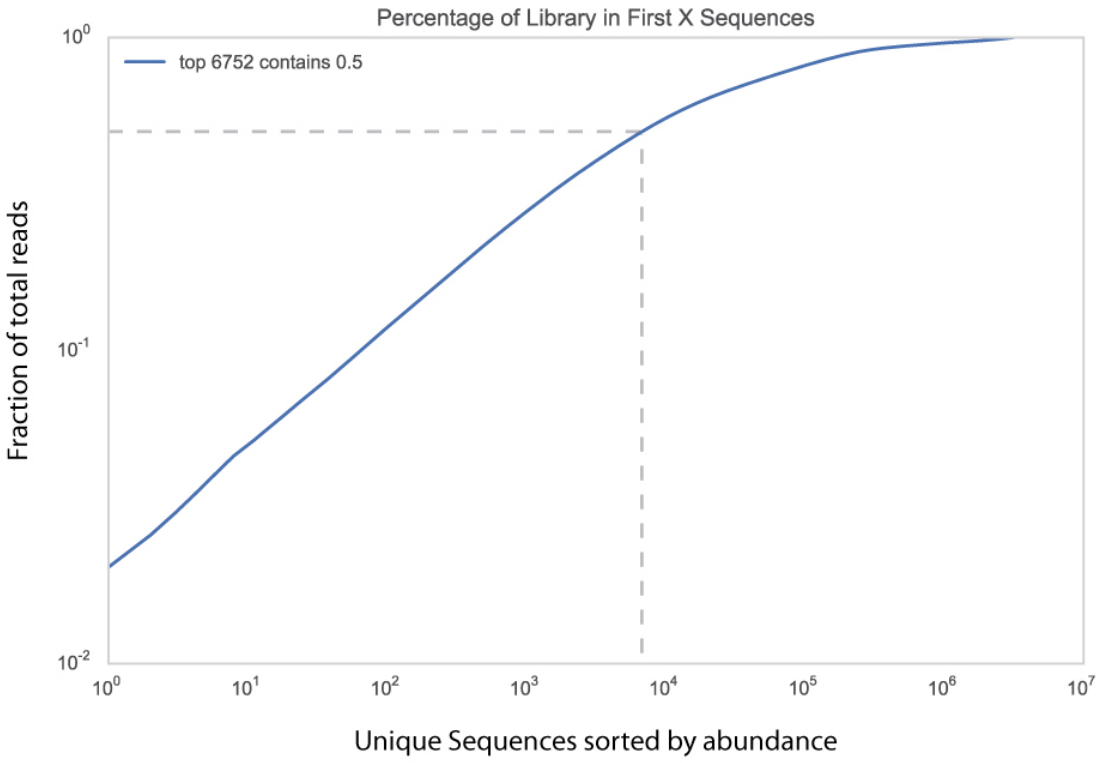


Figure A1-1. Distribution of sequencing reads based on ranked sequence abundance.

To assign a PK structure to each sequence, we chose to assess *compatibility* with PK configurations without accounting for energetic contributions. Because this approach is tailored to our restricted scaffold, it is computationally more efficient than absolute prediction. Notably, this approach often leads to PK assignments that are in agreement with pseudoknot prediction algorithms (such as pKiss). We created a hairpin-type (H-type) PK feature space using the starting library's sequence constraints and our imposed structural constraints. Every H-type PK can be defined by the lengths of its segments: (a) stem 1; (b) loop 1; (c) stem 2; (d) loop 2; and (e) loop 3. To establish the PK position within the library scaffold, we also define the length of the segment of unpaired nucleotides preceding (5' to) the PK, the length of the segment downstream of (3' to) the PK (**Fig. A.1-2**), and fix the total length to 44 nucleotides.

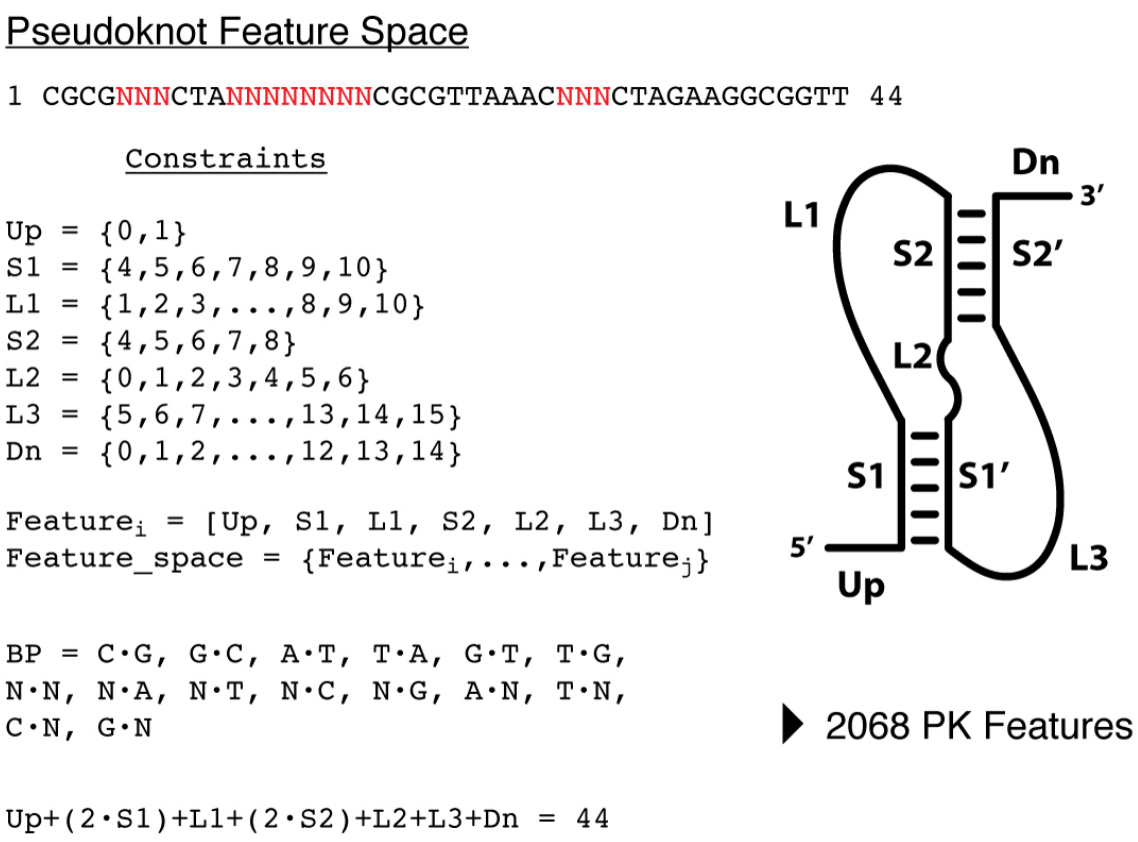


Figure A1-2. Construction of the pseudoknot (PK) feature space.

Using the library scaffold's constraints and our imposed constraints, a total of 2,068 PK features were generated. The top 20,000 most abundant sequences from the NGS, as well as every theoretical starting library sequence, was evaluated for its compatibility with the PK feature space (step 2 of the pipeline). A sequence was deemed compatible with a given PK feature if it supported the base pairs present within the PK feature (allowable base pairs include the standard Watson-Crick A-U, U-A, G-C, and C-G, along with the wobble G·U and U·G pairs). Of the PK features supported by a given sequence, the feature with the most base pairs is chosen as the assigned PK feature for that sequence. In the case of a tie, both PK features are assigned to the sequence. From the theoretical starting library, 1,507 PK features were selected by at least 1 sequence.

In step 3 of the pipeline, we determined the enrichment of each PK feature by calculating its pre-selection probability and comparing it to its post-selection probability (**Fig. A1-3**). Explicitly, pre-selection probability was defined as the total number of sequences from the theoretical starting library that were assigned to the PK feature divided by the total starting sequences (2.68×10^8). Post-selection probability was defined as the total number of sequencing reads that fit to the PK feature (number of unique sequences in that feature multiplied by the mean read-count of sequences within that PK feature) divided by the total sequencing reads (approximately 5.6×10^7). The enrichment factor (EF) is then defined as the ratio of the post-selection probability to the pre-selection probability (**Equation A1-1**).

$$EF = \frac{\left(post-selection\ probability\right)}{\left(pre-selection\ probability\right)} = \frac{(\#\ PK\ reads\ /\ total\ reads)}{(\#\ PK\ initial\ /\ total\ initial)} \qquad \textbf{(Equation A1-1)}$$

Each PK feature was ranked according to its EF, and PK features that showed low occupancy (low total read mass) were removed. Then, PK features ranked within the top

10% of EFs were nominated for primary sequencing clustering analysis. Primary sequence clustering was performed using a greedy algorithm, resulting in families of sequences that comprise the motifs. Clusters containing less than 5 sequences were discarded or ignored. From this analysis, 97 clusters were nominated as final -1 PRF motifs.

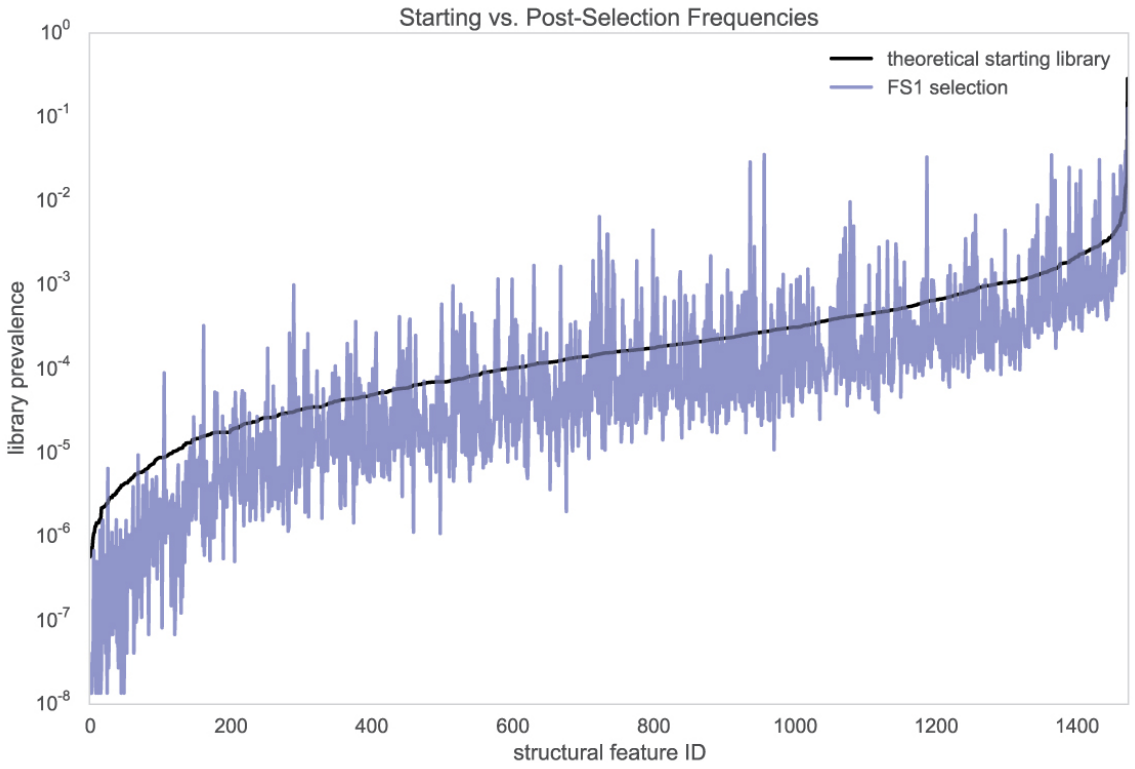


Figure A1-3. PK feature enrichment. The PK feature probability was calculated from the theoretical starting library (assuming equal distribution of sequences) by dividing the number of sequences within a PK feature by the starting library size (approx. 2.68 x 10^8). Post selection probability was computed by dividing the number of sequencing reads contained within the PK feature by the total sequencing reads (approx. 6 x 10^7).

The analysis pipeline generates a summary spreadsheet report for all nominated motifs, including an ID number, the number of sequences supporting the motif, the mean

and standard deviation of abundance level for sequences in the motif, the assigned PK secondary structure, the mode sequence, the neighborhood occupancy about the mode, and the normalized neighborhood entropy about the mode. Additionally, for each motif, the analysis pipeline generates a fasta file containing the motif's sequences and a sequence logo[1] displaying information content (**Fig. A1-4**). An additional variant analysis can be performed as step 4 in the pipeline to analyze the sequence neighborhood of a particular sequence of interest. This output provides a plot of single nucleotide variants (**Fig. A1-5**) for any sequence of interest, as well as a heat map of abundance for variants with two nucleotide (pairwise) differences (**Fig. A1-6**). Notably, the pairwise output may highlight positions where mutual information is present.

Figure A1-4. Sequence logo for -1 PRF motif. Thick bars indicate positions of randomized nucleotides in the starting library.

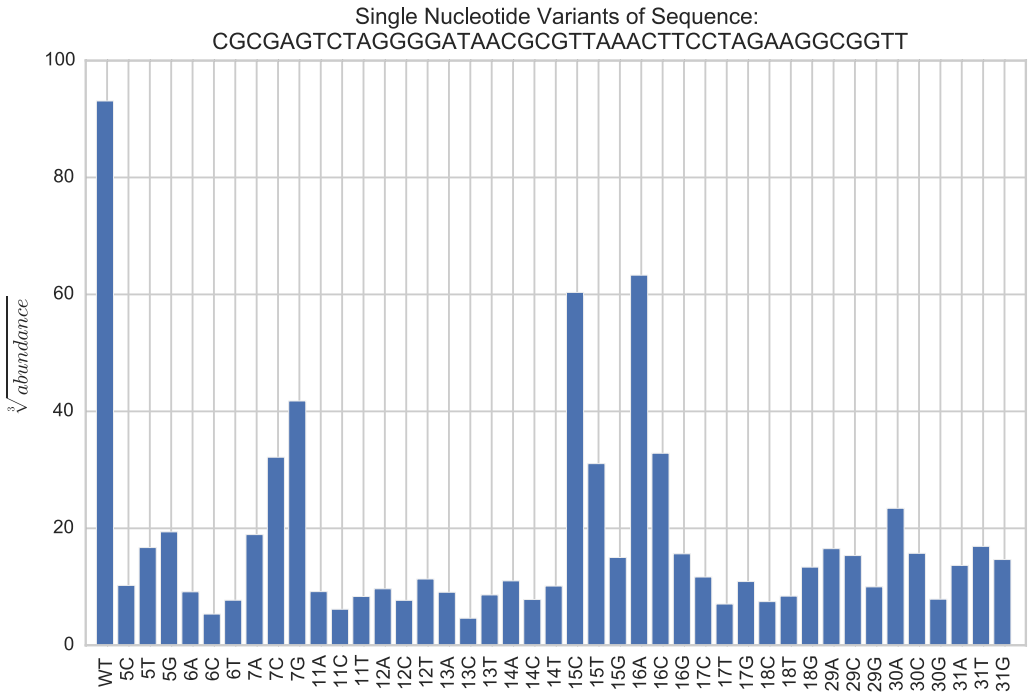


Figure A1-5. Single nucleotide variants of target sequence FS-1.

Figure A1-6. Pairwise variants of the target sequence FS-1.

A2. Thermodynamic calculations for -1 PRF ON-switches.

To design and optimize the ligand responsive ON-switches described in chapter 3.2.2, thermodynamic calculations were performed using RNA secondary structure prediction algorithms. Calculations were performed using NUPACK[2] with Turner parameters[3] and the temperature set to 30 ˚C. By default, NUPACK does not predict pseudoknots and therefore will predict the lowest energy non-pseudoknotted secondary structure for the RNA. pKiss was used to predict FS-2 pseudoknot energy at 30 ˚C. **Table A2-1** summarizes the results of the calculations. RNA segments used for calculations spanned from the start of the FS-2 pseudoknot to the unstructured mid-insulator region.

Ligand binding energy contributions were estimated based on the K_d of the ligand aptamer interaction and the concentration of the ligand that was used. We estimated thermodynamic contributions of RNA folding and ligand binding for 4 hypothetical scenarios: (1) ΔG of the system in the OFF state not bound to ligand; (2) ΔG of the system in the ON state not bound to ligand; (3) ΔG of the system in the OFF state when bound to ligand; and (4) ΔG of the system in the ON state when bound to ligand (**Fig. A-2.1**). The ON state is defined by a folded FS-2 pseudoknot structure. The following energies were calculated as contributors to each of the scenarios:

(1) ΔG of the system in the OFF state not bound to ligand, Σ of:
- Ensemble ΔG of the unfolded pseudoknot
- Ensemble ΔG of the switching hairpin
- Ensemble ΔG of the aptamer region

(Note: Σ of ΔG's can be estimated by the ensemble ΔG of the entire sequence)

(2) ΔG of the system in the ON state not bound to ligand, Σ of:
- ΔG of the folded pseudoknot
- Ensemble ΔG of the sequence 3' to the pseudoknot

(3) ΔG of the system in the OFF state when bound to ligand, Σ of:
- Ensemble ΔG of sequence 5' to the aptamer
- ΔG of the folded aptamer
- ΔG of ligand binding

(4) ΔG of the system in the ON state when bound to ligand, Σ of:
- ΔG of the folded pseudoknot
- ΔG of the folded aptamer
- ΔG of ligand binding

Unbound		*Ligand Bound*	
ΔG OFF (kcal/mol)	**ΔG ON** (kcal/mol)	**ΔG OFF** (kcal/mol)	**ΔG ON** (kcal/mol)
Unfolded FS-2 (ensemble ΔG)	Folded FS-2 (MFE ΔG)	5' of aptamer (ensemble ΔG)	Folded FS-2 (MFE ΔG)
Hairpin (ensemble ΔG)	3' of FS-2 (ensemble ΔG)	Folder Aptamer (MFE ΔG)	Folder Aptamer (MFE ΔG)
3' of Hairpin (ensemble ΔG)		Ligand Binding (ΔG)	Ligand Binding (ΔG)
Σ ΔG Unbound OFF	Σ ΔG Unbound ON	Σ ΔG Ligand OFF	Σ ΔG Ligand ON

Figure A2-1. Description of thermodynamic calculations for evaluation of -1 PRF ON-switches.

The relative calculated energies of the ON and OFF states were used to guide decision for constructing switches, with the desirable property that the OFF state is preferred in the absence of ligand and the ON state is preferred in the presence of ligand (i.e. minimize ΔΔG ON – OFF in the absence of ligand, and maximize ΔΔG ON – OFF in the presence of ligand). When ligand is absent from the system, only scenarios (1) and (2) are considered. However, in the presence of ligand, energies of all 4 scenarios must be considered (ON unbound and ligand bound, OFF unbound and ligand bound). Therefore,

for calculating ΔΔG between the ON and OFF states in the presence of ligand, the Helmholtz free energy was used (-kT ln Q) for the ON and OFF states using the two contributing scenarios (unbound and ligand bound).

Table A2-1. Summary of thermodynamic calculations for -1 PRF ON-switches.

Sequence	Σ ΔG (kcal/mol) Unbound OFF	Σ ΔG (kcal/mol) Unbound ON	Σ ΔG (kcal/mol) Ligand OFF	Σ ΔG (kcal/mol) Ligand ON	ΔΔG (kcal/mol) OFF-ON Unbound	ΔΔG (kcal/mol) OFF-ON Ligand
Theo-ON-1	-32.44	-33.58	-33.16	-36.75	+1.14	+3.43
Theo-ON-2	-34.14	-33.75	-34.86	-36.75	-0.39	+1.73
Theo-ON-3	-35.87	-33.65	-36.59	-36.75	-2.22	+0.00
Theo-ON-4	-35.69	-35.04	-37.35	-38.58	-0.65	+1.19
Theo-ON-5	-36.69	-35.95	-36.77	-39.99	-0.74	+2.84
Theo-ON-6	-36.45	-39.00	-38.18	-43.27	+2.55	+5.05
Neo-ON-1	-25.70	-28.65	-25.71	-32.64	+2.95	+6.51
Neo-ON-2	-27.46	-28.62	-27.55	-32.64	+1.19	+4.73
Neo-ON-3	-31.15	-29.82	-31.02	-32.64	-1.33	+1.14
Neo-ON-4	-29.11	-29.16	-29.02	-32.64	+0.05	+3.16

A.3 References

1. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–6100 (1990).
2. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32,** 170–173 (2011).
3. Serra, M. J. & Turner, D. H. Predicting thermodynamic properties of RNA. *Methods Enzymol.* **259,** 242–261 (1995).